COMPUTATIONAL ANALYSIS OF THE TRANSCRIPTOME USING LONG-READ RNA SEQUENCING

by

Nathan P. Roach

A dissertation submitted to The Johns Hopkins University

in conformity with the requirements for the degree of

Doctor of Philosophy

Baltimore, Maryland

July, 2020

Abstract

Reconstructing the transcriptome from RNA sequencing reads is a challenging problem, especially when no high quality reference genome is available. Current transcriptome annotations have largely relied on short read lengths intrinsic to most widely used high-throughput cDNA sequencing technologies. For example, in the annotation of the *Caenorhabditis elegans* transcriptome, more than half of the transcript isoforms lack full-length support and instead rely on inference from short reads that do not span the full length of the isoform. Short read sequencing technologies, though accurate, cannot reliably reconstruct full-length transcripts due to the highly complex nature of the transcriptome with large gene families, widespread alternative splicing, and highly variable expression and coverage per transcript. We applied nanopore-based direct RNA sequencing to characterize the developmental polyadenylated transcriptome of *C. elegans*. Using this approach we provide support for 23,865 splice isoforms across 14,611 genes, without the need for computational reconstruction of gene models.

In addition, we have developed an open source *de novo* transcriptome assembly method, CONDUIT, which uses single molecule long read RNA sequencing to generate scaffolded splice graphs independent of a reference genome.

It then pseudomaps short-read RNA sequencing reads to isoforms extracted from the scaffolded splice graph, polishes these splice graphs using both short and long read data, and outputs consensus isoforms extracted from these splice graphs. We show that CONDUIT produces highly accurate consensus isoforms, completely independent of a reference genome in several model systems and in a novel pathogenic yeast system.

Primary Reader and Co-Advisor: John Kim

Co-Advisor: James Taylor

Secondary Reader: Michael Schatz

Acknowledgments

John & Anna Roach: I wish you both were still here to see this. I learned so much from you. So much of who I am as a person is thanks to you.

"The Moving Finger writes; and, having writ,
Moves on: nor all your Piety, nor Wit
Shall lure it back to cancel half a line
Nor all your Tears wash out a Word of it."

- Omar Khayyam; *The Rubaiyat*

My advisor James Taylor: I don't know how to express how grateful I am that you welcomed me into your lab. Your advice and feedback was instrumental, your patience as I caught up with where I needed to be was so appreciated. I could not have done this without you. We all miss you.

My advisor John Kim: You were my toughest critic and my fiercest ally. Thank you for your patience, your advice, and for pushing me to be a better scientist. If I had to do it all again, I wouldn't change my decision to join your lab for the world.

Christine Roach: You've always been there for me when I needed you most. I could not have gotten this far without you. I love you Mom.

Timothy Roach: Whenever I was struggling, you listened, and you gave me advice that helped me through. Love you Dad.

Jeremy Roach: I don't think I would have made it without you showing me it was possible. Thank you so much for that.

Marissa Roach: I'm so proud of you Riss. You're doing so great, and I'm so excited to see what you do next. Thank you for all the support, and congratulations again on the Masters!

Carol Demme: Grandma, thank you for everything. I wish I could have made more family dinners while I was away. Hopefully I'll be around more often now. Love you so much.

Vince Demme: Papa, thank you for your support and your love. Can't wait to see you again soon. Love you.

The Gregoires (& Jason): Thank you for the dinners, the O's games, the holidays, and for welcoming me to Maryland. It was so great to have family close by.

The Siskars, Maria Demme, Walter & Marge Kozacki: Thank you for all

the support. I couldn't have done this without my family.

The Taylor Lab: From the moment I joined your lab I felt so welcomed. I could not have made it this far without the support you provided. Thank you.

The Kim Lab: I got through the hardest years of my graduate schooling thanks to the kindness, camaraderie, and determination of the Kim Lab. I hope you all know how appreciated you were.

Jeanna Stokes & Susan Bailey: When I was at my lowest point, your advice, expertise, and counseling got me back on my feet again. I will never forget that. Thank you so much.

Shemika Brooks: Your group was so helpful in enabling me to de-stress, and contextualize my problems. Thank you for everything.

Ryan Clark: You've been my best friend in Maryland, and I don't think I would have made it this far without you being there to grab a drink and de-stress with.

Eric Thompson, Steven Walter & S Muma: The group chat was so helpful when I needed to de-stress. You made me laugh so often. Thank you all.

Ashley Pitt: From our long talks about science in our first year, to checking in on each other as we worked on our research, I couldn't ask for a more supportive, insightful, or funny friend and colleague. Thank you.

Boris Brenerman, Gherman Uritskiy, Max Echterling, Dania Uritskiy & Kevin Macuba: I looked forward to every DnD session, and I'll miss you all so much. Thanks for helping to keep me sane. Village people forever.

My committee Bob Johnston & Mike Schatz: Thank you for the critical feedback, and the excellent advice.

Winston Timp, Norah Sadowski, & Amelia Alessi: This research could not have happened without your contributions. Thanks for the help.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Transcriptome Structure Overview in *Caenorhabditis elegans*

The complete set of RNA products produced by an organism, i.e. the transcriptome, has a profound impact on the physiology and relatedly, the fitness of that organism (Srivastava et al., 2019). Broadly, the transcriptome can be divided into two categories, RNA products that code for protein, or protein-coding RNAs, and RNA products that do not code for protein, or non-coding RNAs (Craig et al., 2014). All protein products produced by the organism must first be encoded as messenger RNA (mRNA), and translated through the action of non-coding ribosomal and transfer RNAs (rRNAs and tRNAs), while regulatory networks of other non-coding RNAs can influence the levels of RNAs themselves, the translation rate of mRNAs, and even the chromatin state of DNA (Craig et al., 2014; Jonas and Izaurralde, 2015; Holoch and Moazed, 2015; Morris and Mattick, 2014).

Our current understanding of the nematode roundworm *Caenorhabditis elegans*

transcriptome has been determined with a variety of techniques including coding sequence (CDS) prediction algorithms, Expressed Sequence Tag (EST) and Open Reading Frame Sequence Tag (OST) cDNA based libraries assessed through Sanger sequencing, and short read based RNA sequencing (collected largely by the modENCODE project but also from other sources) (Spieth et al., 2014; Lamesch et al., 2004; Reboul et al., 2001; Boeck et al., 2016; Hillier et al., 2009). Based on data collected through these and other approaches, the *C. elegans* transcriptome contains approximately 53,000 genes, around 20,000 of which are protein coding (Lee et al., 2018; WormBase web site, 2018).

1.1.1 Protein-coding genes

Generally, protein coding genes consist of a promoter region, a transcription start site (TSS), a five-prime untranslated region (5'UTR), a coding sequence (CDS), a three-prime untranslated region (3'UTR), and a transcription termination region (Craig et al., 2014; Haberle and Stark, 2018; Porrua and Libri, 2015). The 5'UTR, CDS, and 3'UTR are comprised of exonic sequences (regions that are included in the mature mRNA product) and can also include intronic sequences (regions that are removed from the mature mRNA product in a process known as *cis*-splicing) or outronic sequences (regions that are removed from the mature mRNA product in a process known as *trans*-splicing) (Craig et al., 2014). The functional end-products of these genes are the protein product expressed through the translation of the mature mRNA (although secondary functional roles in addition to translation have been identified for several mRNAs) (Craig et al., 2014; Li and Liu, 2019). These mRNA transcripts are often subject to regulation from various protein and ribonucleoprotein (RNP)

complexes, which recognize their binding sites within the mRNA and bind to influence various properties of the transcript including rates of translation and degradation, localization within the cell, RNA base modifications, and poly(A) tail lengths (Craig et al., 2014; Szostak and Gebauer, 2013; Andreassi and Riccio, 2009; Zhao et al., 2015; Eckmann et al., 2011).

As of the WS237 release of the *C. elegans* transcriptome, *C. elegans* coding genes have median length of around 2 kilobases long, though annotated *C. elegans* coding genes range from as short as 30 bp to 102.6 Kb based on the WormBase WS265 transcriptome (A portion of the distribution of which can be seen in Figure 1.1) (Spieth et al., 2014; Lee et al., 2018; WormBase web site, 2018).

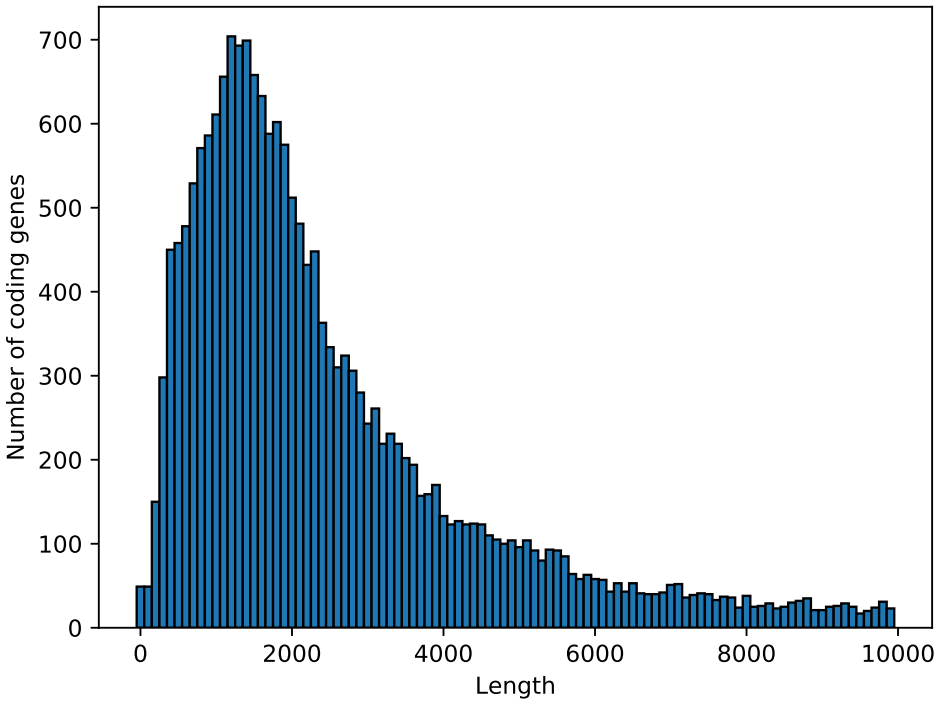


Figure 1.1: Length distribution of WS265 protein coding genes

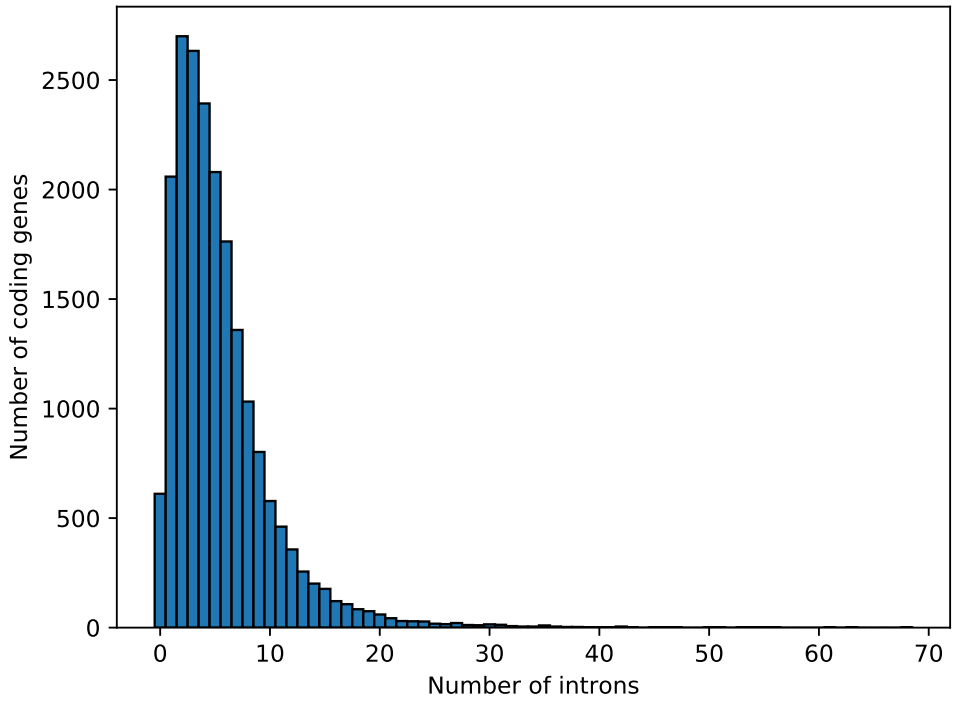


Figure 1.2: Number of introns of WS265 protein coding genes

As of the WS237 release of the *C. elegans* transcriptome annotation, there are 108,151 annotated introns in the protein coding genes of *C. elegans* (Spieth et al., 2014). The number of introns within the pre-mRNA transcript is variable gene to gene, ranging from 0 introns in single exon genes to 68 unique introns in the highly complex gene *ttn-1* (Figure 1.2; based on the WS265 release of the transcriptome annotation) (Lee et al., 2018; WormBase web site, 2018). The length of these introns also varies greatly, ranging from micro-introns approximately 25 nt in length, to the 100 kilobase intron present in the gene *nhr-23* (Spieth et al., 2014). The overall distribution of intron lengths in the WormBase WS265 annotation of the transcriptome can be seen in Figure 1.3

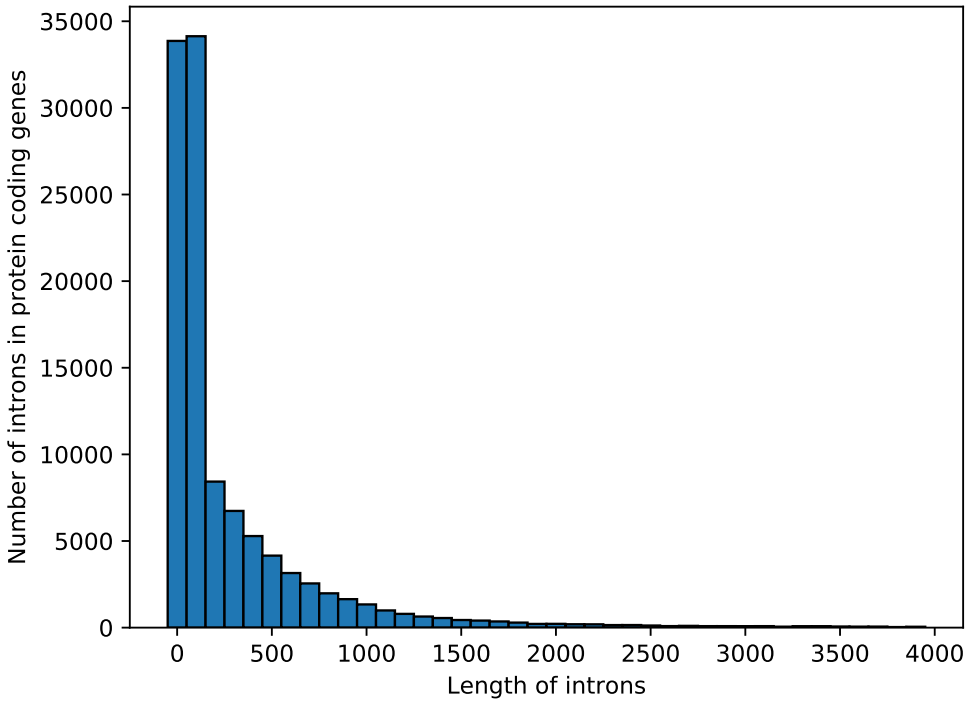


Figure 1.3: Length of introns of WS265 protein coding genes

(WormBase web site 2018; Lee et al. 2018). In general intron lengths in *C. elegans* are much shorter than those in vertebrate organisms, and sequence elements of introns common in vertebrate spliceosomal spliced introns are less common in *C. elegans* (the differences in the prevalence of these sequence elements is discussed in more detail in Section 1.3.2.1) (Riddle et al., 1997; Zahler, 2018).

Similarly, the number of exons per protein coding gene in the annotation is broadly distributed (Figure 1.4), as is the length of these exons (Figure 1.5) (Both based on the WS265 transcriptome release) (Lee et al., 2018; WormBase web site, 2018). Exon length in *C. elegans* is similarly distributed to the exon

lengths of vertebrate organisms (Spieth et al., 2014).

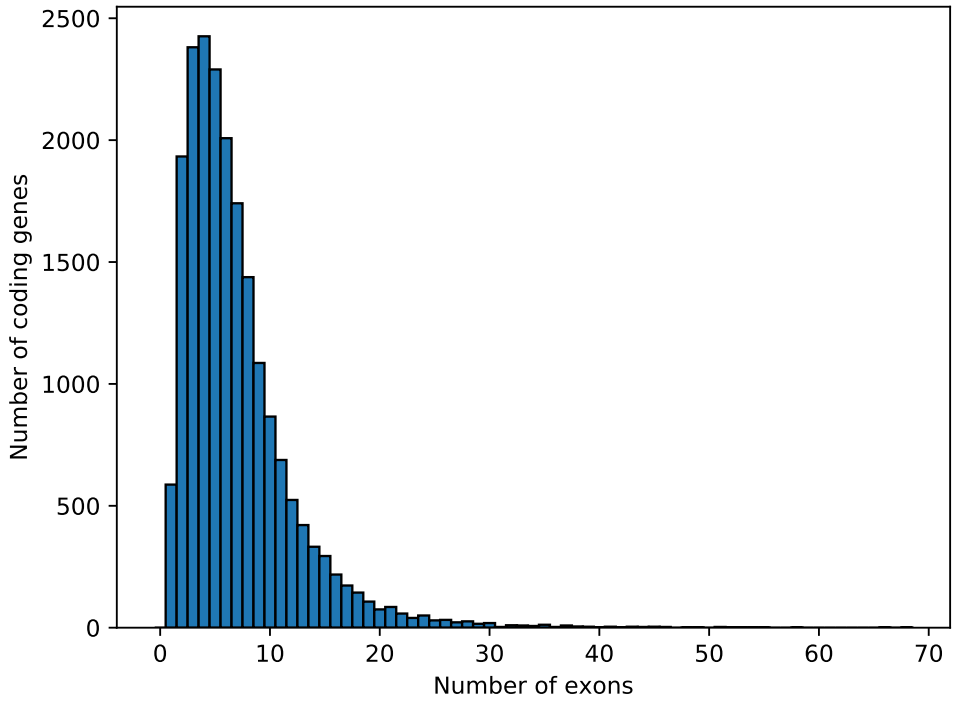


Figure 1.4: Number of exons of WS265 protein coding genes

3'UTRs are important features of protein coding mRNA transcripts that disproportionately contain binding sites for RNA binding proteins and regulatory small RNAs (Szostak and Gebauer, 2013; Andreassi and Riccio, 2009). Many genes can express multiple possible 3'UTRs isoforms, and the usage of longer or shorter 3'UTR isoforms can lead to the imposition of or evasion of additional levels of regulation of an RNA transcript (Mayr and Bartel, 2009). In genes with multiple 3' UTRs shorter 3'UTR isoforms are less likely to have a canonical or alternative poly(A) signal (PAS) specifying the cleavage and polyadenylation site than longer 3'UTR isoforms or 3' UTRs from genes with

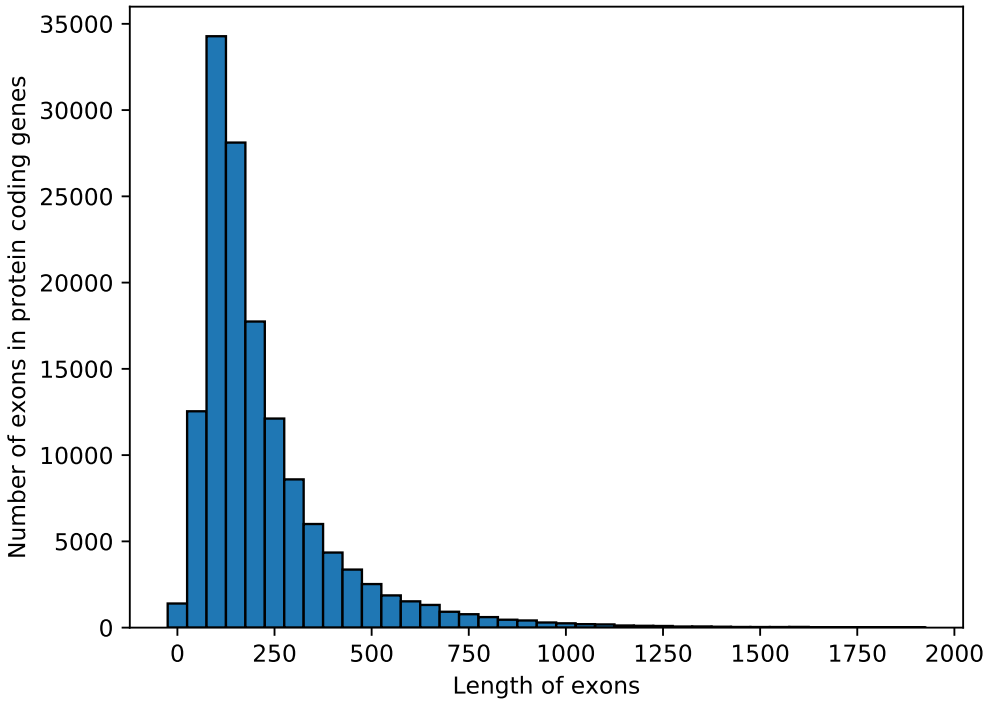


Figure 1.5: Length of exons of WS265 protein coding genes

a single 3′ UTR (Mangone et al., 2010). In *C. elegans* average 3′UTR lengths have been shown to change across developmental stages, getting shorter as the organism gets older (Mangone et al., 2010). In the WS265 annotation of the *C. elegans* transcriptome, many protein coding genes (27.9%) do not have an annotated 3′UTR, despite there being several datasets that have used sequencing to profile 3′UTRs genome wide (Lee et al., 2018; WormBase web site, 2018).

The number of functional transcripts is also dramatically increased in many eukaryotes through the process of alternative splicing, which allows a single gene to encode multiple transcripts (Baralle and Giudice, 2017). The process of

alternative splicing is discussed in more detail below in Section 1.3.2.1.2.

1.1.2 Non-coding RNA genes

Non-coding RNAs serve diverse roles within the cell, and can be broken down into many subclasses. Non-coding RNA loci in *C. elegans* have been discovered and annotated based on a variety of sources, including large scale deep sequencing from various publications (Spieth et al., 2014). We address various subclasses of non-coding RNAs below.

1.1.2.1 Small non-coding RNA genes

Small non-coding RNAs are typically defined as RNA products shorter than an arbitrary threshold of 200 nt in length that do not code for protein (Han Li and Chen, 2015). Most non-coding genes in *C. elegans* are small non-coding RNAs (piRNAs, in particular represent a large fraction of the non-coding RNA in *C. elegans*, 61% of non-coding genes in the WS237 release of the transcriptome annotation) (Spieth et al., 2014). In eukaryotes, many subclasses of small non-coding RNAs are bound by Argonaute proteins and bind their RNA targets through complementary base pairing (depending on the small RNA class this pairing can exhibit perfect complementarity, or have several mismatches) (Hoogstrate et al., 2014; Billi et al., 2014; Vella and Slack, 2005). The effect and physiological role of small ncRNA binding varies by subclass, which will be explained for each subclass in further detail below.

1.1.2.1.1 MicroRNAs (miRNAs)

MicroRNAs are a highly conserved class of small RNAs whose mature product

generally is around 22 nt in length responsible for regulating the stability and protein expression levels of RNA transcripts (Cai et al., 2009). miRNAs are first transcribed as primary RNA (pri-miRNA) transcripts. These transcripts are modified with a 5′ guanosine cap, polyadenylated, and sometimes trans-spliced (Bracht et al., 2004; Lee et al., 2004). These transcripts will have a region or regions in which the RNA will form a double stranded RNA hairpin usually around 60 - 70 nt in length (Vella and Slack, 2005). This hairpin structure is recognized by the protein Drosha, which cleaves the RNA around the hairpin, releasing the hairpin as a pre-miRNA (Lee et al., 2003). This pre-miRNA is exported to the cytosol where it is recognized and further processed by the protein Dicer (Bernstein et al., 2001; Grishok et al., 2001; Vella and Slack, 2005). One strand of this processed RNA is then loaded into an Argonaute protein, which in the case of *C. elegans* miRNAs are named *ALG-1* and *ALG-2* (Grishok et al., 2001).

Once loaded into the Argonaute, the miRNA, the Argonaute, and associated co-factors bind their RNA targets through complementary base pairing of the miRNA to the target, and form the miRNA Induced Silencing Complex, or miRISC (Bartel, 2009; Fabian and Sonenberg, 2012). This complex can have varying effects depending on the cofactors recruited, including acceleration of RNA degradation, silencing of the RNA target through repression of translation, shortening of the poly(A) tail of the RNA target, and in some cases stabilization of the RNA target (Pillai et al., 2005; Eulalio et al., 2008; Dallaire et al., 2018).

miRNAs are highly developmentally regulated, and are known to be involved

in regulating key developmental transitions as well as spatial expression patterns (Johnston and Hobert, 2003; Vella and Slack, 2005). Specifically, in *C. elegans*, the miRNA *lin-4* is known to regulate the transition between the L1 and L2 stages through repression of translation of the *lin-14* mRNA, and the miRNA *let-7* is known to regulate the larval to adult transition through repression of *lin-41* (Olsen and Ambros, 1999; Reinhart et al., 2000). Such roles for miRNAs in regulating developmental timing are broadly conserved across somatic animal development (Ambros, 2011). Though miRNAs have been shown to be regulators of development, the impact of miRNAs in many developmental contexts have not been studied. For example, little is known about the miRNA-ome of worms in and after developing in the dauer state, a stress induced developmental state that results in drastic changes to the developmental program and gene expression patterns.

1.1.2.1.2 21U Piwi interacting RNAs (piRNAs)

Piwi interacting RNAs, or piRNAs are another class of small RNAs, named after the class of argonaute proteins they are known to interact with. The Piwi orthologs of *C. elegans* are *prg-1* and *prg-2*, though *prg-2* is thought to be non-functional as *prg-2* mutants do not result in loss of piRNA expression (Das et al., 2008; Batista et al., 2008). In *C. elegans* piRNAs are consistently 21 nt long and have a 5′ Uridine, and are therefore referred to as 21U piRNAs (Ruby et al., 2006; Das et al., 2008; Batista et al., 2008). Canonically, piRNAs repress the expression of transposable elements in the developing germline, through complementary base pairing to transposable element transcripts, thereby maintaining genomic integrity of this immortal cell lineage (Das

et al., 2008; Malone and Hannon, 2009; Bagijn et al., 2012; Czech et al., 2018). There are many indications that beyond their canonical role of transposon repression in the germline piRNAs also play significant roles in somatic cell lineages and regulate the expression of endogenous mRNA transcripts in both these somatic tissues and the germline (Rouget et al., 2010; Peng and Lin, 2013; Kim et al., 2018a). Mutations of *prg-1* and loss of piRNAs has dramatic impacts on the morphology of the germline and the fertility of affected organisms (Batista et al., 2008). It is unclear, however, whether these morphological and fertility defects are due to de-silenced transposons leading to a destabilized genome or due to dysregulation of endogenous genes in the absence of regulatory piRNAs. 21U piRNAs are capable of triggering production of WAGO class 22G endo-siRNAs, which act as secondary siRNAs to silence gene expression through inhibition of transcription, destabilization of mRNA, and modification of chromatin states of their target genes (Bagijn et al., 2012; Ashe et al., 2012; Shirayama et al., 2012; Luteijn et al., 2012).

Unlike other organisms, piRNAs in *C. elegans* are transcribed individually by RNA Polymerase II and each piRNA represents an independent transcriptional unit (Gu et al., 2012; Billi et al., 2013). These transcriptional units contain upstream sequence motifs that are necessary and sufficient to drive and regulate expression of the piRNA (Ruby et al., 2006; Billi et al., 2013). Given the individually transcribed nature of each piRNA, in the annotation of the *C. elegans* transcriptome each piRNA is annotated as its own gene and as such there are over 15,000 piRNA genes annotated therein (Lee et al., 2018; WormBase web site, 2018).

1.1.2.1.3 Endogenous small interfering RNAs (endo-siRNAs)

Endogenous small interfering RNAs, or endo-siRNAs in *C. elegans* were discovered through deep sequencing of small RNAs and were identified as antisense endo-siRNAs with lengths enriched at 26 and 22 nucleotides (Ambros et al., 2003; Ruby et al., 2006; Han et al., 2009). These endo-siRNAs have a bias for containing a 5' guanosine, and were therefore named 26G and 22G RNAs respectively (Ambros et al., 2003; Ruby et al., 2006; Han et al., 2009; Gu et al., 2009)

26G RNAs

26G RNAs are primary siRNA triggers whose binding to their complementary targets triggers the production of secondary 22G RNAs of the WAGO class (Vasale et al., 2010). 26G RNAs in *C. elegans* broadly fall into one of two classes, the oogenesis enriched 26Gs associated with the argonaute *ERGO-1*, and the spermatogenesis enriched 26Gs associated with the redundant argonautes *ALG-3* and *ALG-4* (Han et al., 2009; Conine et al., 2010). Both sets of 26Gs are produced by the RNA dependent RNA polymerase *RRF-3* and have similar biogenesis machinery, however the production of these pools of 26G RNAs is tightly temporally regulated, and the targets of these two sets of 26Gs are largely distinct (Han et al., 2009; Conine et al., 2010; Billi et al., 2014). The *ALG-3/4* 26G RNAs are primarily produced during spermatogenesis (during the L4 stage of development) and mostly target spermatogenic genes (Conine et al., 2010). As a result loss of *ALG-3/4* 26G RNAs can result in temperature sensitive sterility (Han et al., 2009; Conine et al., 2010). The *ERGO-1* class

26Gs in contrast are produced primarily during oogenesis, however they do not target germline enriched genes and their loss therefore does not result in fertility defects (Vasale et al., 2010; Billi et al., 2014). Though it is not clear what the exact role of *ERGO-1* 26Gs are (as their loss results in an enhanced RNAi phenotype but no other major noted phenotypes), they are highly abundant in the developing oogenic germline, and have been speculated to target newly evolved genes, as their sequence conservation is quite low, even when comparing *C. elegans* to the recently diverged C. briggsae (Yigit et al., 2006; Han et al., 2009; Pavelec et al., 2009; Vasale et al., 2010; Fischer et al., 2011; Billi et al., 2014).

There is some evidence that 26G RNAs originate from spliced template transcripts, as some 26Gs have been shown to span exon-exon junctions (Ruby et al., 2006; Han et al., 2009; Gent et al., 2010). There are multiple lines of evidence that splicing factors play a role in 26G biogenesis. Screens have identified splicing factors as being important for 26G levels, suggesting splicing and 26G pathways are functionally linked or share a common set of regulatory factors (Kim et al., 2005; Robert et al., 2005; Montgomery et al., 2012). The splicing factor *TCER-1*, in particular, was shown to be required for wild type levels of 26G accumulation in at least three 26G RNAs by qRT-PCR (Weiser, 2019). It was also shown that removing the intron of an exogenous gene depleted the levels of a 26G RNA targeting that gene, suggesting that spliceosomal recruitment plays a role in 26G biogenesis (Weiser, 2019).

22G RNAs

22G RNAs are also thought to exist in two classes, namely the WAGO class 22Gs and the *CSR-1* class 22Gs, distinguished by their argonaute cofactors (Billi et al., 2014).

The WAGO class 22Gs are a secondary siRNA product downstream of 26G, exogenous RNAi, or piRNA targeting and are thought to be the major effector of silencing in these pathways (Billi et al., 2014). WAGO 22Gs are highly abundant, and are germline enriched (Billi et al., 2014). These 22G products are produced using the mRNA target of the primary siRNA or piRNA as a template, allowing for the amplification of siRNAs targeting that mRNA (Billi et al., 2014).

CSR-1 22G RNAs in contrast are not thought to be secondary siRNAs and their biogenesis is not dependent on piRNAs or 26G RNAs (Billi et al., 2014). CSR-1 22G RNAs are thought to play a role in gene licensing in addition to gene silencing, and have been shown to regulate chromatin organization (Billi et al., 2014).

1.1.2.2 Long non-coding RNAs (lncRNAs)

Long non-coding RNAs, or lncRNAs, are RNAs longer than 200 nt that do not code for protein, though many lncRNAs are processed in the same manner as pre-mRNAs through 5′ capping, splicing and polyadenylation (Quinn and Chang, 2016; Mercer et al., 2009; Wilusz et al., 2009). Most lncRNAs fall into one of two classes: long intervening ncRNAs (lincRNAs; AKA long intergenic ncRNAs), which do not overlap the exons of protein coding genes, or anti-sense ncRNAs (ancRNAs) which overlap the exons of protein coding genes on

the opposite strand (Nam and Bartel, 2012). Most reported lncRNAs have no known function, however some lncRNAs have been shown to have fundamental roles in gene regulation e.g. the mammalian lncRNA *Xist* has been shown to be required for X Chromosome inactivation (Nam and Bartel, 2012; Borsani et al., 1991; Brockdorff et al., 1992; Brown et al., 1992). lncRNAs can impact gene regulation through several mechanisms, including the recruitment of chromatin modifiers, modulation of alternative mRNA splicing, and acting as a 'sponge' for miRNAs (decreasing the concentration of unbound miRNA capable of repressing it's target genes) (Paraskevopoulou and Hatzigeorgiou, 2016; Marchese et al., 2017). In the WS265 annotation of the *C. elegans* transcriptome, 1,422 lncRNAs are currently annotated, and several papers have sought to expand this annotation through the genome wide identification of additional lncRNAs (Lee et al., 2018; WormBase web site, 2018; Nam and Bartel, 2012; Akay et al., 2019).

1.1.2.3 Transfer RNAs (tRNAs)

Transfer RNAs are RNA products transcribed by RNA polymerase III that are 'charged' with amino acids on their 3' end and are essential to the process of translation as they contribute amino acids to the growing polypeptide chain and act to decode the codons of the mRNA being translated into the appropriate amino acid to be incorporated (Craig et al., 2014). Because of their important role in translation tRNAs are highly expressed and duplicated many times throughout the genome. In the WS265 Wormbase annotation of the *C. elegans* transcriptome, there are 634 annotated tRNAs, and an additional 209 psuedogenic tRNA loci (Lee et al., 2018; WormBase web site, 2018).

1.1.2.4 Ribosomal RNAs (rRNAs)

Ribosomal RNAs (rRNAs) are the most abundant RNA product in any organism and are necessary for the generation of protein products (Craig et al., 2014). Ribosomal RNAs are notable for several reasons including their catalytic activity which drives the formation of peptide bonds in cellular protein synthesis (Craig et al., 2014). Like tRNAs, rRNAs genes are duplicated many times throughout the genome in *C. elegans* (Spieth et al., 2014). The 18S, 5.8S, and 26S rRNA subunits of *C. elegans* are transcribed by RNA polymerase I from a region of Chromosome I thought to contain approximately 55 copies of a tandem repeat of the rDNA unit, while the 5S rRNA is transcribed by RNA polymerase III from a region of Chromosome V thought to contain approximately 110 copies of the 5S rDNA (Spieth et al., 2014; Paule and White, 2000).

1.2 Brief Overview of Development in *Caenorhabditis elegans*

The development of *Caenorhabditis elegans* is highly stereotyped and reliably results in adult hermaphrodites with 959 somatic cell nuclei (or, less commonly, adult males with 1031 somatic cell nuclei) (Alton and Hall, 2009). Development of *C. elegans* starts as a single cell embryo, which for approximately 150 minutes after fertilization develops *in utero*. At around the gastrula stage (~30 cell stage) of embryonic development the eggs are laid, where they develop *ex utero* for approximately 9 hours before eventually hatching into the first larval stage of *C. elegans* development, the L1 stage (Alton and Hall, 2009).

In continuous development, the worm then develops through the next three larval stages, L2, L3, and L4 (Alton and Hall, 2009). During each of these stages worms express stage specific proteins and develop a stage specific cuticle (each of which is molted during the transition from one stage to the next) (Cox and Hirsh, 1985; Hillier et al., 2009). After the L4 stage the worm undergoes one final molt, shedding its L4 cuticle and emerging as a young adult with a functional germline (Alton and Hall, 2009). Because most worms are hermaphrodites, they are capable of self fertilizing, and will begin to lay eggs approximately 8 hours later, repeating this process (Alton and Hall, 2009)

There have been many studies examining the transcriptome of *C. elegans* in various developmental contexts using short read sequencing, focused on 1) characterizing the genes expressed in the developmental stages and several time points in embryonic development (Boeck et al., 2016; Hillier et al., 2009; Gerstein et al., 2010), 2) the 3'UTR choices in each developmental stage (Mangone et al., 2010), 3) the genes, isoforms and 3'UTRs expressed in the various sections of the germline arm (West et al., 2018; Diag et al., 2018), 4) the transcriptome of an adult worm at single cell resolution, and more (Packer et al., 2019).

In addition to continuous hermaphroditic development, there are a number of alternative developmental pathways that *C. elegans* can proceed down depending on environmental and genetic factors. Although most worms will develop as hermaphrodites (specified through the inheritance of two X Chromosomes), a small subset of worms (< 0.5% in wild-type populations)

will develop as males due to chromosomal nondisjunction events resulting in the inheritance of only one X Chromosome (Hodgkin et al., 1979). Males express different genes and undergo different developmental pathways than hermaphrodites and as such exhibit different behavioural and morphological phenotypes (Albritton et al., 2014).

There are also several stress induced alternative developmental pathways that can occur in the course of *C. elegans* development. In the absence of sufficient food L1 worms can transition to an L1 diapause (L1d) developmental state in which the worm ceases development temporarily until food becomes available again (Baugh, 2013). After the L1 state worms also have the ability to transition to L2 pre-dauer state (L2d), and will do so based on various environmental factors including high temperature, crowding or starvation (Karp, 2018; Alton and Hall, 2009). These L2d worms will then eventually develop into dauer worms, a stress resistant state in which worms develop thick cuticles containing alae, seal off their oral orifices, constrict their pharynx, and slow their metabolism (Alton and Hall, 2009). Worms can persist in this stress resistant state for up to 4 months, and upon dauer exit will resume normal development as post-dauer L4 worms (Alton and Hall, 2009). Post-dauer L4 worms are morphologically indistinguishable to normal L4 worms, however they express different genes than L4 worms resulting from continuous development, indicating that there are lasting impacts of the dauer state on the transcriptome (Dalley and Golomb, 1992; Wang and Kim, 2003; Karp et al., 2011; Hall et al., 2010, 2013).

1.3 Processing of RNAs in *Caenorhabditis elegans*

1.3.1 Capping

7 methylguanosine (m7G) 'caps' are a prevalent modification of RNA products in eukaryotic organisms (Craig et al., 2014). All eukaryotic mRNA transcripts are co-transcriptionally capped through a 5′ to 5′ triphosphate linkage between the m7G cap and the 5′ most base of the transcript (Craig et al., 2014; Ramanathan et al., 2016). This linkage is produced through the action of several enzymes. First, a 5′ phosphate is dephosphorylated through the action of a phosphatase, following this a guanosine monophosphate is joined to the resulting diphosphate forming the 5′ to 5′ linkage (Craig et al., 2014; Ramanathan et al., 2016). In *C. elegans* both of these steps are catalyzed by the same bifunctional enzyme: *CEL-1* (Takagi et al., 2003). The 5′ G cap is then methylated through the action of a methylase, which in *C. elegans* is thought to be *TAG-72* based on homology to other guanine N7 methyltransferases (Craig et al., 2014; WormBase web site, 2018; Lee et al., 2018; Shaye and Greenwald, 2011; Kim et al., 2018b).

The resulting m7G cap protects the mRNA from the action of 5′ to 3′ exonucleases (Furuichi et al., 1977; Shimotohno et al., 1977; Cowling, 2009). In addition to its role in protecting the mRNA from degradation, the presence of a 5′ cap recruits various 5′ cap binding complexes, which inhibit decapping enzymes and, when in complexes with poly(A) binding protein, act to promote translation (Schwartz and Parker, 2000; von der Haar et al., 2004; Kahvejian et al., 2005; Li and Kiledjian, 2010).

1.3.2 Splicing

1.3.2.1 *cis*-splicing

Most protein-coding genes (96.9%), and some classes of ncRNAs in *C. elegans* undergo RNA splicing in *cis*, a series of biochemical events (namely two trans-esterfication reactions) in which precursor mRNA (pre-mRNA) transcripts are processed to remove intron sequences and join exon sequences together (Craig et al., 2014; Harris et al., 2010; Lee et al., 2018; Tourasse et al., 2017).

1.3.2.1.1 Spliceosomal driven splicing in *C. elegans*

Although in some organisms some introns are self splicing, most splicing in *C. elegans* (and eukaryotes in general) relies on the action of the spliceosome, a molecular complex comprised of various proteins and small nuclear ribonucleoproteins (snRNPs) that act to promote splicing through a series of biochemical steps (Craig et al., 2014; Riddle et al., 1997; Zahler, 2018). The five snRNPs that comprise the majority of the spliceosome are simply named U1, U2, U4, U5 and U6 (Craig et al., 2014). These components are highly conserved, particularly the sequence regions that interact with and recognize sequences that define splice sites (Indeed these sequence regions are perfectly conserved between vertebrates and *C. elegans* in the U1 and U5 snRNPs) (Riddle et al., 1997). In addition to these core snRNP components several proteins are also necessary for spliceosome function, including the branchpoint binding protein BBP (AKA splicing factor 1 (*SF-1*)), and the U2 auxiliary factor proteins *U2AF65* and *U2AF35*, which together form the U2AF complex (in *C. elegans* these genes are named *sfa-1*, *uaf-1*, and *uaf-2* respectively) (Zahler, 2018;

Zorio et al., 1997; Zorio and Blumenthal, 1999b; Mazroui et al., 1999).

Spliceosomal driven splicing follows a biochemical process that involves the formation of a 5′ to 2′ linkage forming a lariat structure between a GU sequence just 3′ of the 5′ splice site and an A nt downstream (Craig et al., 2014). This lariat formation frees the hydroxyl at the 3′ end of the upstream exon, which then attacks the phosphate linking the 3′ of the intron to the 5′ of the downstream exon, thereby splicing the two exons together, and removing the intron sequence (Craig et al., 2014). In spliceosomal driven splicing, these reactions are facilitated (and possibly directly catalyzed by) the action of the snRNP and protein components of the spliceosome (Craig et al., 2014).

Spliceosomal driven splicing, however, requires sequence components to drive binding and action of the various spliceosomal subcomponents. Namely, in most eukaryotes (including *C. elegans*) the boundary of most introns are defined at their 5′ end with a GU sequence, and at their 3′ with an AG sequence (Riddle et al., 1997). Interestingly, in *C. elegans*, this 3′ AG sequence is part of a larger, well conserved 3′ splice site sequence UUUUCAG|R not found in most other eukaryotes (Riddle et al., 1997; Zahler, 2018). Most eukaryotes contain a polypyrimidine tract sequence between the branch point A nt and the 3′ splice site (excluding *C. elegans*, which has no obvious enrichment for such sequence tracts beyond the short polypyrimidine tract present in it's consensus 3′ splice site) (Riddle et al., 1997; Zahler, 2018). The eukaryotic polypyrimidine tract is known to be the binding site for the U2AF complex (an essential component of the spliceosome) in most organisms, making the absence of such tracts in *C. elegans* notable (Craig et al., 2014; Riddle et al., 1997; Zahler, 2018). It has

been shown, however, that both components of the U2AF complex bind the extended UUUUCAG|R 3′ splice site of *C. elegans*, suggesting this extended 3′ splice site is responsible for the recruitment of U2AF in *C. elegans* mRNAs (Zorio and Blumenthal, 1999a). In addition to sequences that recruit the U2AF complex, the recognition of a putative branchpoint for lariat formation by branchpoint binding protein / splicing factor 1 (BBP / *SF-1*) family proteins is necessary for spliceosome assembly in other organisms (Craig et al., 2014; Liu et al., 2001; Selenko et al., 2003). However, unlike other eukaryotes, there is no strong consensus sequence for branch points identified in *C. elegans*, drawing into question how the *C. elegans* BBP / *SF-1* homolog *SFA-1* is recruited to precipitate spliceosome assembly (Zahler, 2018). Binding studies performed *in vitro* show that *SFA-1* strongly binds to human *U2AF65*, suggesting that *SFA-1* may be recruited to branch points *in vivo* through interactions with U2AF protein subunits (Zahler, 2018; Hollins et al., 2005).

The steps of spliceosomal splicing have largely been discovered in other organisms and from *in vitro* studies (Cheng and Abelson, 1987; Konarska and Sharp, 1986; Ruby, 1997; Seraphin and Rosbash, 1989; Pikielny et al., 1986). Given the short intron length of *C. elegans*, the lack of a conserved branchpoint sequence motif, and the longer UUUUCAG|R conserved 3′ splice site of *C. elegans*, it is possible that the order of spliceosomal component recruitment and the drivers of that recruitment may be different in *C. elegans* than in other organisms. That said, in *in vitro* systems derived from yeast or HeLa cells' spliceosomal components the first step of spliceosomal splicing is the binding of the U1 snRNP to the 5′ splice site, the binding of BBP to the branchpoint

sequence, and the binding of the U2AF complex to the 3′ splice site (Brow, 2002; Craig et al., 2014). The U2 snRNP is then recruited to the branchpoint and takes the place of the BBP (Brow, 2002; Craig et al., 2014). U4 U5 and U6 snRNPs are then recruited, a conformational change occurs releasing the U1 and U4 snRNPs, and the first transesterification reaction occurs releasing the 5′ exon and forming the lariat at the branch point (Craig et al., 2014). Additional conformational rearrangements occur bringing the splice sites in proximity and catalyzing the second transesterification reaction joining the two exons and releasing the lariat intron (Craig et al., 2014).

1.3.2.1.2 Alternative splicing

Many genes in *C. elegans* encode for multiple transcripts due to alternative splicing, in which transcripts originating from the same gene are differentially spliced, resulting in different exons being incorporated (or not incorporated) into the final RNA product (Zahler, 2005; Lee et al., 2018; WormBase web site, 2018). This differential splicing can take many forms, including utilization of alternative 5′ or 3′ splice sites, the inclusion or exclusion of cassette exons (exons that can either be spliced into the final transcript or not), the splicing of mutually exclusive exons (sets of two or more exons in which one and only one of the exons can be incorporated into the final transcript), and intron retention (in which a sequence that can be spliced out as an intron is instead retained in the mature mRNA transcript) (Zahler, 2005). In addition to alternative splicing, alternative polyadenylation (APA) sites or alternative transcription start sites (TSS) can also result in multiple isoforms at the same gene (See Figure 1.6 for an overview of the types of alternative transcript structures)

(Figure inspired by Zahler 2005) (Zahler, 2005). The number of *C. elegans* genes with a given number of isoforms in the WormBase WS265 annotation is displayed in Figure 1.7 (Lee et al., 2018; WormBase web site, 2018). As can be seen, most genes in *C. elegans* encode for only one functional transcript, and are not alternatively spliced, and the number of genes that express a given number of isoforms monotonically decreases as the number of isoforms increases. This is not the case in humans and mice, where most genes express two or more alternative isoforms (Bussotti et al., 2016; Djebali et al., 2012; Pan et al., 2008; Wang et al., 2008).

The mechanisms by which alternative splicing is regulated in *C. elegans* has been discussed in several review articles (Zahler, 2018; Gracida et al., 2016; Wani and Kuroyanagi, 2017). Briefly, alternative splicing is dictated by the assembly location of spliceosomes on the pre-mRNA transcript (Zahler, 2018). At alternatively spliced splice junctions the locations the spliceosome assembles is thought to be driven by the presence of trans-acting splice factors which can recruit or inhibit spliceosome assembly (Zahler, 2018). These splice factors are in turn recruited to the pre-mRNA transcript by *cis*-acting sequence elements of the pre-mRNA (Wang and Burge, 2008; Zahler, 2018). The prescence, or lack thereof of functionally active *trans*-acting factors therefore must drive alternative splicing patterns, as the *cis*-acting elements of the pre-mRNA are constant between cell types (Zahler, 2018). Supporting the model of *cis*-acting elements driving alternative splicing regulation is the fact that primary sequence is sufficient to predict cryptic splicing patterns and the emergence of new alternative splice sites with reasonable accuracy using deep learning

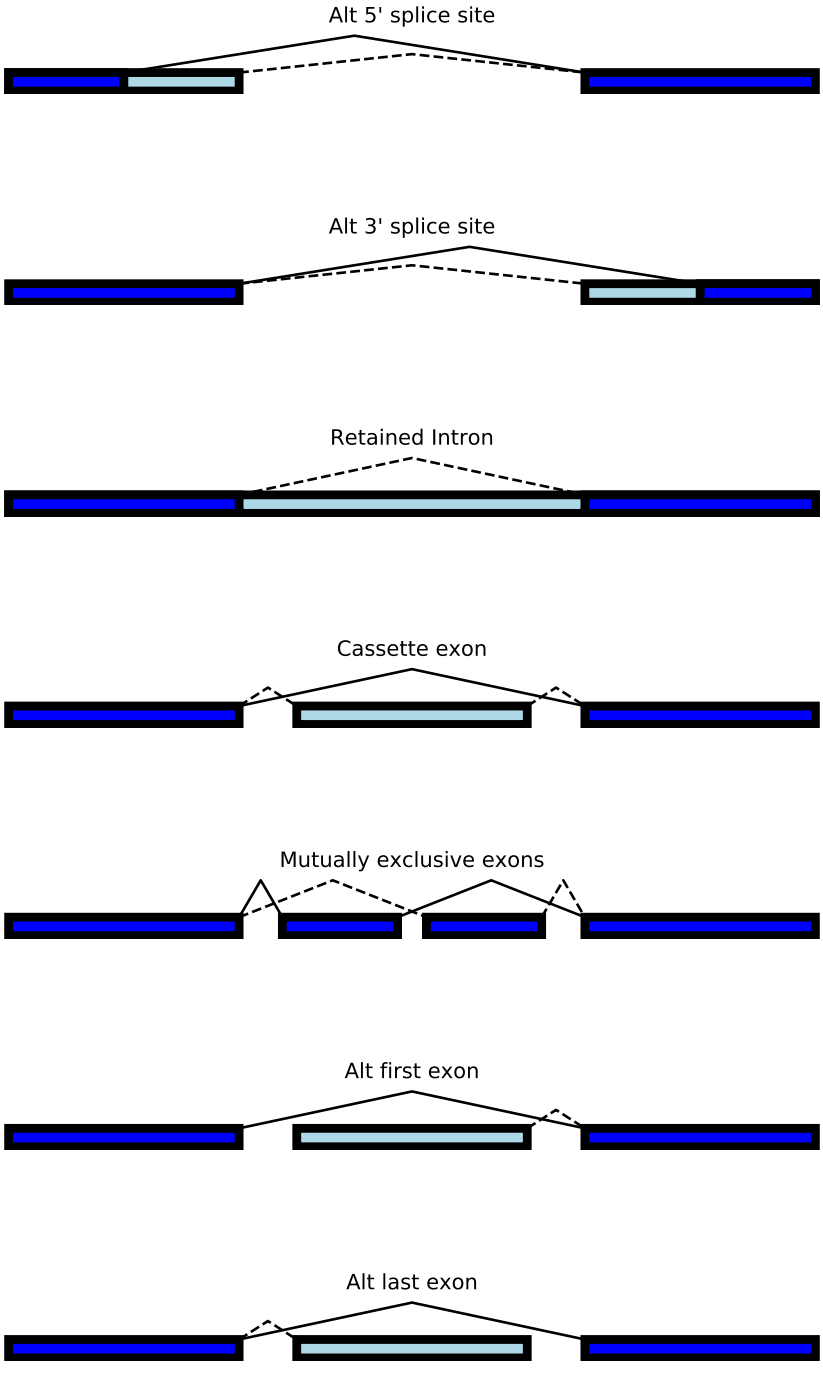


Figure 1.6: Types of alternative transcript structures

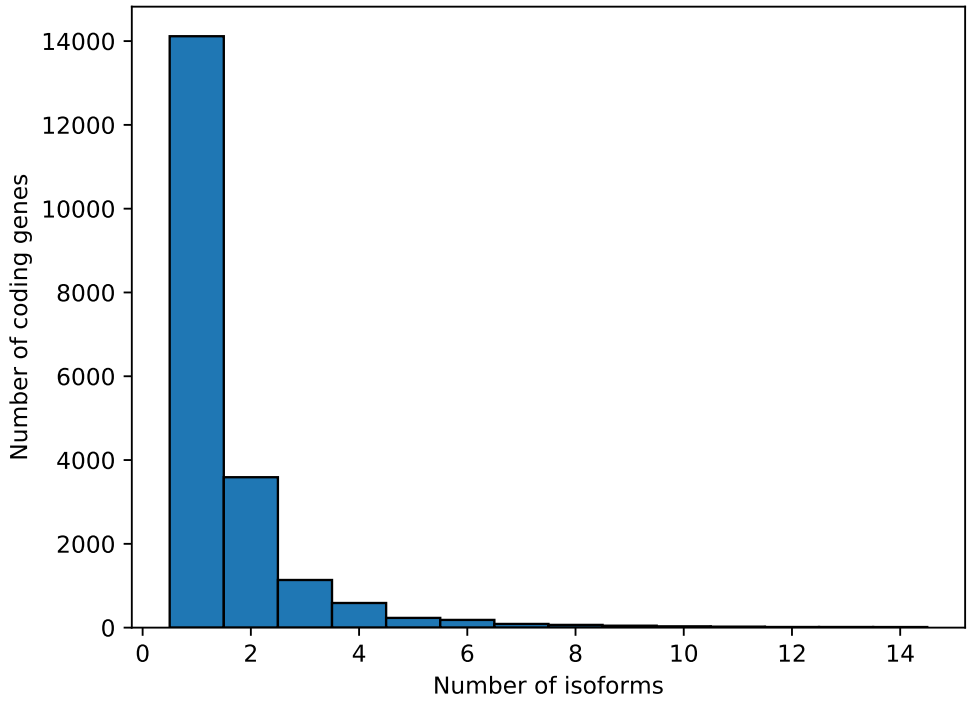


Figure 1.7: Number of isoforms of WS265 protein coding genes

models trained on pre-mRNA sequence and corresponding RNA-sequencing data (Jaganathan et al., 2019). Many of these *cis*-acting elements and their putative trans-acting binding partners have been catalogued, however additional factors likely remain to be discovered (Calarco et al., 2011; Zahler, 2018). Two broad classes of these trans-acting splicing factors include SR (Ser-Arg) family proteins and hnRNPs (heterogeneous nuclear ribonucleoproteins) (Gracida et al., 2016). In *C. elegans*, SR family proteins are called *rsp* genes, and based on RNAi experiments appear to be partially functionally redundant (Kawano et al., 2000; Longman et al., 2000, 2001; Zahler, 2018). Generally, SR proteins

are thought to promote splicing through the recruitment of spliceosomal components, while hnRNPs are thought to inhibit splicing, however exceptions to this generalization have been identified (Gracida et al., 2016; Cáceres et al., 1994; Smith and Valcárcel, 2000; Huelga et al., 2012).

Though accurate prediction of alternative splicing patterns remains a difficult and largely unsolved problem, the regulation of alternative splicing in specific *C. elegans* genes has been elucidated. One notable such example is the splicing of the gene *unc-52*, which in the presence of functional *mec-8* preferentially skips exons 17 and 18 (Lundquist et al., 1996; Zahler, 2018). Viable mutations in exons 17 and 18 of *unc-52* are synthetically lethal with mutations in the *mec-8* gene (Lundquist and Herman, 1994; Zahler, 2018). This synthetic lethality has been used to identify additional splicing regulators (*smu* genes) through a genetic screen (Lundquist and Herman, 1994; Zahler, 2018). Another example is the splicing of *egl-15* exons 5A and 5B. Exon 5B incorporation is inhibited by the action of *asd-1* (a *fox-1* family gene) / *fox-1* and *sup-12* the binding motifs of which have been identified (Gracida et al., 2016; Kuroyanagi et al., 2006, 2007). The structural mechanism of repression has been proposed to involve the action of the RNA-recognition motifs of both *ASD-1* and *SUP-12* sterically inhibiting the binding of constitutive splicing factors to regions of intron 4, though the exact mechanism remains unknown (Amrane et al., 2014; Kuwasako et al., 2014; Gracida et al., 2016). Complicating prediction of splicing regulation is the fact that some splicing factors have been shown to have context dependent regulation, such that in some cases the factor inhibits splicing, and in other cases the factor promotes splicing (Motta-Mena et al.,

2010; Gracida et al., 2016). *ASD-1 / FOX-1* is one such example, inhibiting incorporation of exon 5B in *egl-15* as mentioned above, but also promoting incorporation of exon 7A in *unc-32* in neurons (Kuroyanagi et al., 2013; Gracida et al., 2016).

Alternative splicing has also been shown to generate isoforms which are differentially targeted by the nonsense mediated decay (NMD) pathway (Hansen et al., 2009; Muir et al., 2018). Many of the targets of this AS-NMD pathway are themselves splicing factors, suggesting a feedback loop linking AS and NMD pathways that has been supported by studies in several systems (Morrison et al., 1997; Wollerton et al., 2004; Ni et al., 2007). This AS-NMD pathway has been shown to be developmentally regulated in *C. elegans*, leading to differential isoform expression ratios between developmental stages due to differences in the rate of NMD rather than differences in transcription rates of each isoform (Barberan-Soler et al., 2009). This study also suggested that some stabilized transcripts containing a premature termination codon are capable of being translated based on the presence of these transcripts in the polysome fraction of a sucrose gradient experiment (Barberan-Soler et al., 2009).

1.3.2.2 *trans*-splicing

An alternative form of splicing, *trans*-splicing (in which splice leader RNAs are spliced into the 5′ end of an RNA transcript thereby removing outron sequences), is also quite common in *C. elegans* (Allen et al., 2011; Zorio et al., 1994). Splicing in *trans* has been observed to occur in around 70% of genes in the *C. elegans* genome, and is thought to occur with high efficiency in these genes (in other words, if a transcript is *trans* spliced, almost all identical

transcripts will be trans spliced) (Allen et al., 2011; Tourasse et al., 2017; Blumenthal, 2005).

Splicing in *trans* occurs through a similar biochemical mechanism to splicing in *cis*, in which a 5′ to 2′ linkage occurs between the 5′ end of the outron of the SL sequence and the 2′ hydroxyl of a branchpoint in the outron of the pre-mRNA forming a Y like structure (Blumenthal and Thomas, 1988). The freed 3′ end of the SL sequence then attacks a 3′ splice site in the pre-mRNA forming a 5 to 3′ linkage between SL RNA and mRNA, and removing the outron sequence of the pre-mRNA (Blumenthal and Thomas, 1988). It is currently not understood how the SL RNA sequence recognizes a 3′ splice site only when there is no upstream corresponding 5′ splice site in the transcript (Blumenthal, 2005).

Although many splice leader sequences in *C. elegans* have been identified, there are two main splice leader RNAs in *C. elegans* that account for the majority of *trans* splicing that occurs in the organism (Ross et al., 1995; MacMorris et al., 2007). These splice leader RNAs, SL1 and SL2, are both 22 nt in length, and are donated by a roughly 100nt snRNA containing the SL sequence (Ross et al., 1995; Blumenthal, 2005). SL1 class splice leaders are typically donated to transcripts that contain sequences similar to the 3′ splice sites of introns at their 5′ end and are recognized by the SL1 snRNP complex (Blumenthal, 2005; Allen et al., 2011). SL2 sequences however are spliced primarily into *trans* splice sites between transcripts expressed together as a poly-cistronic pre-mRNA from *C. elegans* operons (Blumenthal et al., 2002; Blumenthal, 2005). SL1 *trans* splicing has also been noted to separate genes from poly-cistronic

sequences, however in these instances the *trans* splice site for the SL1 sequence is typically directly adjacent to the cleavage and polyadenylation site of the upstream transcript, unlike SL2 *trans* splice sites, which are typically separated from their upstream genes by approximately 100 nt (Blumenthal, 2005).

There is evidence that *trans*-splicing serves to enhance translational efficiency, suggesting a role for *trans* splicing beyond simply removing unnecessary 5′ sequences and breaking up polycistronic pre-mRNAs (Yang et al., 2017).

1.3.3 Polyadenylation

Polyadenylation is another important and highly prevalent RNA modification in eukaryotes (Craig et al., 2014). Polyadenylation is a chemical modification in which the 3′ end of an RNA transcript is cleaved and a tail of adenosines is added to the cleavage site (Craig et al., 2014).

In most RNA transcripts, there is at least one sequence motif upstream of the cleavage and polyadenylation site that signals the approximate location for the transcript to be polyadenylated (Mangone et al., 2010). This motif is canonically AAUAAA, however similar motifs have been shown to be capable of driving polyadenylation, and it should be noted that polyadenylation has been shown to occur in some locations that contain no such sequence motif (Fitzgerald and Shenk, 1981; Mangone et al., 2010). Canonically, this upstream sequence recruits a cleavage and polyadenylation specificity factor (*CPSF*; by homology in *C. elegans*, the genes *cpsf-1*, *cpsf-2*, *cpsf-3*, and *cpsf-4*) which in complex with a number of additional proteins, cleaves the 3′ end of the mRNA transcript, typically after a CA dinucleotide motif (Proudfoot, 2011; Lee et al.,

2018; WormBase web site, 2018; Shaye and Greenwald, 2011; Kim et al., 2018b). Following this cleavage, untemplated adenosines are added to the newly free 3′ end of the transcript through the action of a poly(A) polymerase (*PAP*; which based on homology in *C. elegans* is produced from the genes *pap-1*, *pap-2*, and *pap-3*) (Craig et al., 2014; Eckmann et al., 2011; Shaye and Greenwald, 2011; Kim et al., 2018b; Lee et al., 2018; WormBase web site, 2018). As the tail is being extended nuclear poly(A) binding protein (*PABPN*; *pabp-2* in *C. elegans*) is recruited to the tail (Eckmann et al., 2011). There is evidence that *PABPN* binding itself regulates the length of poly(A) tail deposition through regulating the interaction between the *PAP* and *CPSF* (Eckmann et al., 2011; Murthy and Manley, 1995). This poly(A) addition is thought to add a poly(A) tail up to 200 adenosines in length (Eckmann et al., 2011). There is ample evidence however, through sequencing studies, that most poly(A) tails *in vivo* exist at lengths significantly shorter than the approximately 200 nt lengths the tails are thought to be deposited at (Chang et al., 2014; Lim et al., 2016; Lima et al., 2017; Subtelny et al., 2014).

Poly(A) tails are thought to promote translation and protect transcripts from degradation, as when poly(A) tails get too short (typically less than approximately 30 nt) the RNA is less likely to be translated and is more likely to be degraded (Nudel et al., 1976; Preiss et al., 1998; Atwater et al., 1990). However, highly expressed genes have been shown to have shorter poly(A) tail lengths on average according to several studies, and ribosome profiling correlates poorly with poly(A) tail length in yeast, several human cell lines, and gastrulation stage embryos of *Danio rerio* and *Xenopus laevis*, suggesting the

relationship between poly(A) tails and transcript regulation may be more complicated than the simple model in which shorter poly(A) tails are more likely to be degraded and less likely to be translated (Subtelny et al., 2014; Lima et al., 2017; Legnini et al., 2019).

1.4 Overview of sequencing approaches utilized in characterizing the *C. elegans* transcriptome

1.4.1 Expressed Sequence Tags (ESTs)

Expressed sequence tags, or ESTs, are a traditional technique used to characterize transcriptomes. In this approach, mRNA is extracted, and usually pulled down by it's poly(A) tails (Parkinson and Blaxter, 2009). This RNA is then reverse transcribed into cDNA, inserted into a plasmid, and these plasmids are transformed into competent bacterial cells. Random bacterial clones are then selected, the plasmids from these clones are extracted, and the cDNA clone is then sequenced in a single-pass sequencing read using Sanger sequencing (Parkinson and Blaxter, 2009). Sequencing can be performed from one or both ends of the cDNA insert.

Much of the early annotation of the *C. elegans* transcriptome was supported through the use of ESTs, and this EST data is still used in the generation of the WormBase transcriptome annotation today (Note that although the *C. elegans* EST library produced by Yuji Kohara was and is widely used, the data was never published on directly, hence the lack of primary citation of this dataset) (Spieth et al., 2014; Mangone et al., 2010).

1.4.2 Open-reading-frame sequence tags (OSTs)

OST based approaches are quite similar to EST based approaches, but after generation of cDNA include a PCR amplification step in which putative open reading frames (ORFs) are amplified precisely from initiation codon to termination codon prior to insertion into bacterial vectors (Walhout et al., 2000; Reboul et al., 2001). OST based approaches therefore require prior knowledge of the putative ORFs one wishes to sequence, as PCR primers must be designed for each of these ORFs. Once this library of ORF insert clones is produced, the resulting bacterial vectors are sequenced using Sanger sequencing (Walhout et al., 2000). These sequences are then aligned to the genome, used to determine the ORFs that were sequenced and to create a curated list of ORFs, the ORF-eome.

1.4.3 Illumina Sequencing

Illumina based sequencing is a short read sequencing approach that sequences DNA molecules by observing the incorporation of fluorescently labeled nucleotides as a complementary DNA molecule is synthesised (Stark et al., 2019; Mardis, 2008, 2013). The DNA molecules to be sequenced are first ligated with adapter sequences which will hybridize to oligonucleotides on the flow cell. Polymerases then produce complementary sequences of the hybridized DNA molecules to be sequenced, and the original DNA molecule is then washed off of the flow cell. The DNA molecules to be sequenced are amplified using bridge PCR, resulting in clusters of clonal molecules, and the sequencing reaction is performed (Mardis, 2008, 2013). In this reaction, fluorescently labeled

nucleotides with a chain terminator are applied to the flow cell, and allowed to be incorporated into an elongating DNA molecule complementary to the template sequence. Each nucleotide has a different fluorophore attached, and the fluorescent signal at each cluster is observed using a microscope. The fluorophores and chain terminators are then cleaved off, and the next base is then incorporated. This is repeated until reads of the desired length are obtained. Once the first sequencing reaction is complete, there is an optional paired sequencing reaction, in which the opposite strand of the DNA molecule is sequenced in the opposite direction (Mardis, 2008, 2013). Since Illumina sequencing is based on sequencing clusters of clonal molecules through observation of the incorporation of complementary bases, errors can occur when the sequencing reactions in the cluster become out of sync with one another (Buermans and den Dunnen, 2014). The percentage of out of sync molecules per cluster increases as the number of elongation cycles increase, and as a result the error rate increases as a function of read length. The maximum useful read length from Illumina sequencers is therefore approximately 300nt, however most Illumina experiments use read lengths substantially shorter than this (Buermans and den Dunnen, 2014).

RNA-seq on Illumina platforms relies on first reverse transcribing RNA molecules into cDNA, and then sequencing the cDNA (Stark et al., 2019). Extensive amounts of Illumina RNA-seq data of *C. elegans* have been produced, and have been used to guide and improve the WormBase annotation of the transcriptome (Hillier et al., 2009; Spieth et al., 2014; Boeck et al., 2016).

In addition to its utility in RNA sequencing, through specialized library preparation approaches and analysis, Illumina sequencing has been used to profile poly(A) tail lengths, alternative polyadenylation sites, and alternative transcriptional start sites in *C. elegans* (Mangone et al., 2010; Jan et al., 2011; Saito et al., 2013; Lima et al., 2017).

1.4.3.1 Illumina based poly(A) tail profiling (TAIL-seq / mTAILseq / PAL-seq)

The challenge of estimating the length of the poly(A) tail using Illumina sequencing is non-trivial, in part because Illumina error rates increase as a function of read length, and poly(A) tail lengths can range in the hundreds of basepairs, and in part because the error rates also increase quite dramatically when sequencing homopolymers (Luo et al., 2012). That said, several biochemical and computational approaches to sequencing poly(A) tails have been developed, including PAL-seq, TAIL-seq, and it's successor mTAIL-seq. TAIL-seq is performed by depleting total RNA of rRNA, ligating biotinylated 3' adapters to the resulting RNA, partially digesting the RNA with RNase T1, pulling down the 3' ligated product with streptavidin, size selecting the fragments by gel purification, 5' adapter ligation, and finally reverse transcription, PCR amplification, and paired end sequencing (Chang et al., 2014). mTAIL-seq operates almost identically, but at the 3' adapter ligation step uses a splint oligo dT adapter to reduce the amount of required input RNA and the rates of internal priming (Lim et al., 2016). In both TAIL-seq and mTAIL-seq Read 1, which ideally lies in the 3' UTR of the transcript, is sequenced for 51 nt, and subsequently mapped and used to determine the transcript of origin

of the poly(A) tail identified (Chang et al., 2014; Lim et al., 2016). Read 2, which measures the poly(A) tail, is sequenced for 251 nucleotides, and the relative amount of raw fluorescent signal that can be attributed to Thymidine incorporation is used as the input to a hidden Markov model (HMM) trained to estimate the number of bases in the poly(A) tail (Chang et al., 2014). This HMM is trained on the relative T signal from RNA spike-ins with poly(A) tails of known length. The estimation of poly(A) tail length for each read is determined by the Viterbi algorithm decoding of the hidden states of the model for that read.

The initial molecular biology involved in preparing samples for PAL-seq is almost identical to that of mTAIL-seq, and also involves splint oligo ligation for ligation of the 3' sequencing adapter, RNase T1 digestion, streptavidin pulldown, etc (Subtelny et al., 2014). PAL-seq differs from mTAIL-seq in the sequencing step itself. Rather than utilizing standard Illumina sequencing approaches, PAL-seq instead flows in a mix of dTTP and a low concentration of biotin-dUTP (Subtelny et al., 2014). Though incorporation of biotin dUTP is random, the high number of clonal molecules in the Illumina clusters results in biotin-dUTP incorporation per cluster that scales highly linearly with the length of the poly(A) tail in the clonal molecules of that cluster (Subtelny et al., 2014). By flowing in fluorescently tagged streptavidin molecules the fluorescence intensity can be measured, and the length of the poly(A) tail in a given cluster can be estimated (Subtelny et al., 2014). The poly(A)-proximal region of the cDNA is sequenced as normal, allowing for the association of poly(A) tail lengths to individual genes.

1.4.3.2 Illumina-based 3'UTR profiling in *C. elegans* (3P-Seq)

Determining the 3' most base of RNA transcripts before cleavage and polyadenylation is necessary for determining 3'UTRs, an essential step in annotating the transcriptome and an important annotation feature for decoding regulation of any given protein coding gene. Poly(A)-position profiling by sequencing (3P-Seq) is one molecular biology technique that has been used in *C. elegans* for profiling 3'UTR structures using short read sequencing technologies (Jan et al., 2011). 3P-Seq starts similar to PAL-seq, with ligation of a biotinylated splint 3' oligo, an RNaseT1 digestion, and a streptavidin pulldown (Jan et al., 2011). Following this, a primer is annealed to the 3' oligo and a reverse transcription reaction is performed with only dTTP present such that only the poly(A) tail is reverse transcribed (Jan et al., 2011). An RNase H reaction is performed, which releases the polyadenylated 3' ends of the RNA transcript from the biotinylated primer (as RNase H preferentially cleaves RNA in DNA / RNA hybrids) (Jan et al., 2011; Hausen and Stein, 1970; Cerritelli and Crouch, 2009). These polyadenylated ends are then purified and prepared for sequencing using standard sequencing approaches (Jan et al., 2011). This allows for sequencing of the terminal 3' end of RNA transcripts, allowing for the precise cleavage and polyadenylation sites of genes to be identified computationally, which is discussed further in Section 1.5.9.

1.4.3.3 Illumina-based TSS profiling in *C. elegans* (5' SAGE / Pol II Initiation Site profiling)

Identifying the TSS of genes in *C. elegans* is made more complicated by the high prevalence of *trans* splicing (which occurs in approximately 70% of *C. elegans*

genes) (Allen et al., 2011). Determining TSS in *C. elegans* therefore requires specialized sequencing protocols to ensure that the 5′ ends of transcripts identified are true TSS and not the 5′ most base before a *trans* splice site. Two papers, Saito et al. and Chen et al., have profiled TSS through 5′ SAGE and a protocol for identification of RNA Polymerase II initiation sites in *C. elegans* respectively (Saito et al., 2013; Chen et al., 2013).

Saito et al. performed 5′ SAGE on RNA extracted from nuclei purified from embryos and adult tissues (grown at 16C to slow the rate of splicing) (Saito et al., 2013). 5′ SAGE involves removal of the 5′ guanosine cap following BAP and TAP treatment, ligation of a 5′ Illumina RNA linker, 1st strand cDNA synthesis via a random primer, 2nd strand synthesis using a primer complementary to the Illumina RNA linker, followed by gel fractionation, size selection, and Illumina sequencing (Hashimoto et al., 2004; Kasai et al., 2005). The resulting data is strongly enriched for reads at the 5′ end of the RNA transcript, allowing for the determination of TSS sites through computational techniques that are described below in Section 1.5.10.

Chen et al. also extracted RNA from purified nuclei, however Chen et al. prepared long and short cap RNA-seq libraries. Preparing short cap RNA-seq libraries involves extracting total nuclear RNA, running this RNA on a polyacrylamide gel and size selecting for fragments between 20 - 100 nt long (Chen et al., 2013). This extracted RNA is then treated with a 5′ to 3′ exonuclease (to remove fragments without a 5′ cap), cloned, PCR amplified with Illumina RNA-seq adapters, and sequenced (Chen et al., 2013).

Long cap RNA-seq libraries similarly starts with total nuclear RNA. It is

then treated with DNAse I to remove genomic DNA contamination, selected for RNA longer than 200 bp using spin columns, and treated with a 5′ to 3′ exonuclease (again to enrich for molecules with 5′ caps) (Chen et al., 2013). Libraries are then constructed using a dUTP replacement method to determine strandedness, and the resulting library is then sequenced (Levin et al., 2010; Chen et al., 2013). The long and short cap libraries were then analyzed computationally to identify putative Pol II initiation sites through computational methods described below in Section 1.5.10.

1.4.4 Pacific Biosciences (PacBio) sequencing

Pacific Biosciences (PacBio) sequencing is a single molecule sequencing technique that involves observing in real time the incorporation of fluorescently labeled nucleotides into a molecule of DNA (Rhoads and Au, 2015). The technology works by ensuring one and only one molecule of DNA is present in each zero mode waveguide (ZMW), an approximately 70 nm wide hole in the single molecule real time (SMRT) cell (Rhoads and Au, 2015; Bleidorn, 2016). This SMRT cell is illuminated from below by an excitatory beam, and the ZMWs are designed such that the intensity of this illumination drops precipitously as the distance from the bottom of the ZMW increases ensuring that only the bottom 20 - 30 nm of the ZMW are illuminated (Rhoads and Au, 2015; Bleidorn, 2016). A complex of DNA polymerase and the DNA molecule to be sequenced is immobilized to the bottom of the ZMW, and fluorescently labeled nucleotides are introduced to the ZMW chamber. When these nucleotides are present in the detection volume of the ZMW their fluorophores are excited and they emit a pulse of light at a wavelength specific to

each nucleotide. When the nucleotide is incorporated the phosphate group is cleaved releasing the fluorophore. By continuously detecting the fluorescence signal of the detection volume, the nucleotide sequence of the DNA molecule can be determined. The error rates of this technology can range between 10 - 15%, however by circularizing the DNA molecule, and sequencing the same molecule many times over, a circular consensus sequence (CCS) for that molecule can be determined, substantially reducing the error rate (Rhoads and Au, 2015; Bleidorn, 2016). In addition, the read lengths possible from PacBio SMRT sequencing are orders of magnitude longer than Illumina sequencing, and typically range from tens to hundreds of Kb long.

A modified approach to PacBio SMRT sequencing called FLAMseq has been used to characterize properties of the full-length *C. elegans* transcriptome in the L4 and young adult stages, including isoform expression levels, poly(A) tail lengths, and 3'UTR lengths (Legnini et al., 2019).

1.4.5 Nanopore sequencing

Nanopore sequencing contrasts sharply with the sequencing approaches discussed above as nanopore sequencing determines sequence by measuring properties of the DNA or RNA molecule directly, rather than measuring the incorporation of complementary nucleotides to the molecule (Stark et al., 2019; Bleidorn, 2016). Nanopore sequencing requires a pore embedded in a non-conductive membrane (with sensors such that the current across the membrane can be measured), and a library of molecules to be sequenced prepared with adapter sequences (bound to motor proteins that will control

the rate of molecule translocation through the pore) (Kono and Arakawa, 2019). A constant voltage is applied across the membrane, thereby driving the negatively charged DNA or RNA molecules through the pore (Kono and Arakawa, 2019). As the single stranded DNA or RNA passes through the pore, the bases resting within the pore at any given time exhibit a given electrical resistance determined by their chemical structure (Maitra et al., 2012; Kono and Arakawa, 2019). Given the constant voltage these variable resistances result in changes in the current across the membrane, which can be measured several thousands of times per second as the molecule passes through the pore (Lu et al., 2016; Kono and Arakawa, 2019). The resulting raw current signal (often referred to as a squiggle) is then passed through trained machine learning models that predict and report the set of bases that best explain the amperages observed (in a process known as basecalling) (Rang et al., 2018). As this approach does not rely on sequencing by synthesis and operates on single molecules (rather than clonal clusters that can get out of sync with one another), its error rates don't increase as a function of the read length, meaning there is no theoretical upper limit to the read lengths obtainable from this technology, and nanopore sequencing reads as long as 2.3 megabases in length have been reported (Payne et al., 2019; Stark et al., 2019). There are limitations to this technology however, notably, nanopore sequencing's error rates are quite high (typically between 5 - 15%, dependent on the versions of the pore, buffer chemistry, and basecalling software used), and these errors are predominantly insertions and deletions (Rang et al., 2018; Wick et al., 2019). In addition, nanopore sequencing cannot capture the 5' most 10 - 15 bases of each sequenced molecule, due to limitations in the motor protein pore

complex assembly (Workman et al., 2019). These properties (high error rates and truncated 5′ sequencing) make nanopore sequencing a poor choice for sequencing short DNA or RNA molecules.

One additional advantage of nanopore sequencing is that it relies only on the size and shape of the molecule passing through the pore, and not on any extension chemistry or complementary base pairing. As a result, RNA can be sequenced directly without the need to reverse transcribe the RNA into cDNA, and modifications to the RNA or DNA molecule such as m6A or 5mC methylation marks result in changes to amperage signals and can therefore be called directly during basecalling (Stark et al., 2019; Stoiber et al., 2017; Ni et al., 2019; Simpson et al., 2017).

Basecalling in the context of nanopore sequencing is the process by which continuous raw amperage signal is converted to a sequence of discrete nucleotide calls. There have been many basecallers released by independent research groups and Oxford Nanopore Technologies, each with distinct machine learning models (Rang et al., 2018; Wick et al., 2019). Each of these models rely on a process known as training, in which reads of known sequence and measured amperage are used to estimate the underlying parameters of the model.

Early ONT models for basecalling relied on Hidden Markov Models (HMMs) which were trained and evaluated using discrete 'event' calls, in which a preprocessing algorithm was used to determine when kmer to kmer transitions were occuring, and segmented the raw signal into these discrete 'events' (Rang et al., 2018). These models rely on the fact that at any given time, the majority of the variability of amperage across the early (R9.4) pores is determined by

only approximately 4-6 bases, allowing for the mean and variance of amperage signal for a given kmer to be calculated and used as learned parameters of the model (Rang et al., 2018). The space of possible kmer to kmer transitions is restricted in this model based on the fact that for any given consecutive kmers the suffix of the first kmer must be the prefix of the second (Wick et al., 2019; Rang et al., 2018). This allowed for reasonably accurate, and fairly fast basecalling.

The next basecalling models released by ONT also relied on discrete event calls, but instead utilized a Recurrent Neural Network (RNN) for their underlying model structure. RNNs are a type of neural network particularly useful for classifying sequences of data (Lipton et al., 2015). In these models, the ith data point in a sequence of data receives information about previous data points through connections from the hidden layers of the (i-1)th data point to the hidden layers of the ith data point (Lipton et al., 2015). This information can then influence the classification output of the model for the ith data point. The information to be passed is learned by the model, which can learn and represent highly complex and non-linear influences of the previous data on the data point to be classified (Lipton et al., 2015). Since the information is passed from one event to the next, and since the weights of transitions from previous hidden layers to subsequent hidden layers are learned, information can be passed recursively, and information from the first event can influence the classification of the final event, depending on the underlying structure of the data (Lipton et al., 2015). This model outperformed the HMM based basecallers in per read and consensus accuracy (Wick et al., 2019; Rang et al.,

2018).

The next advancement in basecalling models was the introduction of a transducer to the RNN model. This model no longer assigned a k-mer to each event call from the raw signal, and instead used the model signal and information from previous events to determine whether to emit no bases, one base, or multiple bases (Rang et al., 2018). (A review that discusses the various iterations of the ONT basecallers indicates that the information passed from previous data points were the bases emitted by previous iterations, however RNNs are typically structured such that they transmit information from the hidden layers, not the classification they emit (Rang et al., 2018; Lipton et al., 2015). As the underlying software is unfortunately not open source, it is difficult to determine whether they are correct in their description of the transducer RNN model).

Next, ONT transitioned away from using event detection in their basecalling approaches, instead training the RNN based model on raw amperage signal (Rang et al., 2018). This model also uses the transducer structure described above.

The most recent versions of ONT basecalling rely on a 'flip-flop' model of basecalling (Wick et al., 2019). In this model bases are again called individually rather than as a set of kmers. The model incorporates two types of states for each base that can be predicted, a flip (+) and flop (-) state (Oxford Nanopore Technologies, 2019; Brown, 2019). This allows the model to distinguish between amperage signal staying at the same level due to a lack of translocation through the pore, and the signal staying at the same level due to

homopolymer stretches (Oxford Nanopore Technologies, 2019; Brown, 2019). This is possible since in decoding steps the model alternates between (+) and (-) bases in homopolymer stretches (Brown, 2019). This results in a significant increase in basecalling accuracy (Wick et al., 2019).

1.4.6 Additional approaches for annotating the 3′ UTRome in *C. elegans*

In addition to 3P-Seq, several other sequencing approaches have been used to annotate the *C. elegans* 3′UTRome including several that have been discussed above. Poly(A) capture followed by Roche 454 sequencing, targeted 3′ RACE using several sequencing methods, Sanger sequencing of cDNA, and Illumina RNA-seq datasets were combined into a 3′UTRome dataset by Mangone et al and used to annotate the 3′ UTRs of the *C. elegans* transcriptome (Mangone et al., 2010).

The poly(A) capture method utilized by Mangone et al. involved poly(A) selection of mRNA by oligo(dT) magnetic beads, followed by reverse transcription with a biotinylated reverse primer and the second strand cDNA was synthesized. This cDNA was then digested with DpnII, the 3′ ends of the transcripts were pulled down with streptavidin, 5′ adapters were ligated to these cDNAs, and the resulting fragments were sequenced using pyrosequencing (Mangone et al., 2010).

The targeted 3′ Rapid Amplification of cDNA Ends (3′ RACE) method used by Mangone et al. involved RNA extraction from mixed staged worms followed by a reverse transcription reaction and a PCR reaction in which one primer

non-specifically targets the poly(A) tail of the resulting cDNA. The other primers included are gene specific, and in the case of Mangone et al targeted the 3'UTRs of 7,105 CDSs across 6,741 genes identified through assays of the ORFeome and Promoterome (Lamesch et al., 2004; Dupuy et al., 2004; Mangone et al., 2010). The resulting 3' ends were then cloned into Gateway vectors and collected as minipools, one for each target which were sequenced by Sanger sequencing. Colonies from these minipools were then combined into eight barcoded libraries, which were then sequenced by Illumina and 454 pyrosequencing platforms (Mangone et al., 2010).

The resulting sequencing data from the various techniques described above were pooled into a 3'UTRome dataset, which Mangone et al used to annotate putative 3'UTR cleavage and polyadenylation sites through computational techniques that will be described below in Section 1.5.9.

1.5 Existing computational tools for annotating transcriptomes

1.5.1 Coding Sequence (CDS) / Open Reading Frame (ORF) prediction algorithms

The ability to predict regions of the genome that code for protein based on genomic sequence alone was extremely important in early attempts to annotate the *C. elegans* transcriptome. The *ab initio* CDS prediction program GeneFinder in particular was used in early WormBase annotation of the transcriptome (Green and Hillier, unpublished software) (Hillier et al., 2005). Though the specific implementation behind the Genefinder program was

never published on, the structure of CDS / ORF prediction algorithms at the time often relied on Hidden Markov Models (HMMs) to predict likely CDS regions (Mathé et al., 2002). HMMs are a type of statistical model in which the process being modeled is assumed to be approximated by a Markov chain (Rabiner, 1989). A Markov process is a process in which there are a fixed set of possible states and a set of possible emissions from each state, the probability of an emission occurring is dependent only on the state, and the probability of being in a state is dependent only on the previous state (Rabiner, 1989). In gene prediction HMMs, the various features of genes to be predicted are included as possible states in the model, and after the parameters of the model were estimated, Viterbi decoding is run on the genomic sequence to determine the most likely sequence of states in the model, and thereby determine the regions of the genome that most likely correspond to coding sequences.

1.5.2 Cufflinks

Cufflinks is a transcriptome assembler designed for short reads that operates on the principle of parsimony to determine the expressed transcripts in a reference genome guided assembly of the transcriptome (Trapnell et al., 2012). Cufflinks operates by first grouping reads at a given locus together into an overlap graph, a data structure in which sequenced fragments are nodes, and nodes are connected if they have compatible splice patterns and they overlap (Trapnell et al., 2012). Once this overlap graph is constructed, the minimum path cover through the graph is calculated (Trapnell et al., 2012). The minimum path cover is a concept from computational graph theory, and represents the smallest set of unique paths through a graph such that every node in the graph

is visited at least once (Cormen et al., 2009). These paths through the graph are then treated as the putative transcripts, and the estimated expression level of these transcripts is then determined in a maximum likelihood based approach (Trapnell et al., 2012).

Though a parsimony based approach for transcript abundance estimation worked fairly well, there are several problems with the approach utilized by Cufflinks, including the fact that it does not consider transcript abundance when determining transcript structures, which can result in an incorrect set of isoforms being called when operating with short reads (Trapnell et al., 2012; Pertea et al., 2015, 2016). Like all short read assemblers, Cufflinks relies on computational inference *post hoc* to determine transcript structures, and therefore often made mistakes when determining transcripts.

1.5.3 StringTie2

StringTie2 is the second major release of the transcriptome assembler StringTie (Pertea et al., 2015; Kovaka et al., 2019). StringTie is a short read reference guided transcriptome assembler that simultaneously builds reference transcriptomes and estimates transcript coverage of those transcriptomes (Pertea et al., 2015). Briefly, StringTie requires short reads aligned to the reference genome using a spliced aligner (Pertea et al., 2015; Kovaka et al., 2019). These alignments are then clustered together based on overlap to produce a splice graph representation of the loci to be annotated (Pertea et al., 2015). Once the splice graph is produced, the 'heaviest path' through the graph is identified (Pertea et al., 2015). The heaviest path is identified in the following manner:

1) the position in the splice graph with the highest coverage is determined and used as the starting point for the path, 2) The path is extended along the graph in a greedy manner, following edges that represent the highest coverage through the graph consistent with the path walked so far (i.e. if a splice junction is present in a read and not present in the heaviest path walked up to that point, that read will not contribute to determining the heaviest path in subsequent extensions), 3) the coverage for this path through the graph is estimated by solving a maximum-flow problem determining the maximum number of reads that can be associated with the transcript. 4) the weights of the splice graph are updated to remove reads that were assigned to a transcript by the algorithm 5) steps 1 - 4 are repeated until all reads are assigned to a transcript or the coverage of the heaviest path drops below some fixed threshold (2.5 reads per bp by default in Stringtie 1.0) (Pertea et al., 2015).

StringTie2 relies on much of the core implementation of StringTie, but incorporated several key improvements. The first major improvement was including support for long read RNAseq such as ONT sequencing or PacBio sequencing (Kovaka et al., 2019). The inclusion of such long, error prone data required key tweaks to the underlying Stringtie algorithm, as errors in the alignment of these long reads at splice sites will lead to incorrect edges in the assembled splice graph. StringTie2 handles this possibility in several ways. The first is a step included in most long read transcriptome annotation tools: correction of splice sites that appear to be wrong. StringTie2 handles this correction step by checking all splice sites present in the alignments high-error reads (usually ONT or non-CCS PacBio reads) (Kovaka et al., 2019). If the splice

site in question is not supported by a low-error alignment read then StringTie looks for a splice site within 10 bp to identify the splice site supported by the most alignments, and adjusts the read's alignment to match this splice site (Kovaka et al., 2019). The next step for dealing with errors in the alignments of ONT and PacBio sequencing was the addition of a pruning step that reduces the number of nodes in a splice graph below some fixed size before evaluation (this size is selected as a parameter and is set by default to 1000) (Kovaka et al., 2019). This pruning algorithm works by removing the edges in the graph with the lowest amount of support from the considered reads (Kovaka et al., 2019). Once these edges are removed the algorithm removes vertices if they are no longer connected to the graph, and merges vertices representing adjacent genomic positions if there is no intron separating them (Kovaka et al., 2019). This improves StringTie2's efficiency, at the expense of possibly pruning out real splice isoforms present in low abundance from the splice graph (typically isoforms from the low throughput long read RNA-seq) (Kovaka et al., 2019).

An additional major improvement was a more comprehensive implementation for incorporating 'super-reads' assembled from input short reads (Kovaka et al., 2019). Super-reads are a type of synthetic long read assembled from short reads by extending the ends of short reads until there is no longer a unique extension to the super read end (Zimin et al., 2013). In the initial StringTie release, super-reads were used only to fill in the gaps between paired-end reads (Pertea et al., 2015; Kovaka et al., 2019). However, in StringTie2 super-reads are extended to their full possible length and used directly in

the assembly of transcript annotations (Kovaka et al., 2019). These changes improved the sensitivity and specificity of short read transcript discovery in this assembly algorithm (Kovaka et al., 2019).

1.5.4 FLAIR

FLAIR, or Full-Length Alternative Isoform analysis of RNA, is a long read analysis pipeline originally designed for use with native nanopore direct RNA sequencing, however this pipeline can also be used with nanopore cDNA sequencing, as well as PacBio sequencing (Tang et al., 2018). This pipeline first involves aligning reads to a reference genome, which is encapsulated in a submodule called flair-align, which aligns reads to the genome using the long read aligner minimap2, and converts the output to a more easily parsed file format (Tang et al., 2018; Brooks, 2019). Reads are then passed to a submodule flair-correct, which is used to correct the splice junctions present in the reads (Tang et al., 2018; Brooks, 2019). These splice junctions are "corrected" in that the program assumes that non-canonical splice junctions not present in a matched short read sequencing sample or a transcriptome annotation are not valid, and in these instances shift these splice junctions to correspond to short read or annotation derived splice sites within 10 nt of the original splice site position (Tang et al., 2018). Once reads are corrected they are passed to the submodule flair-collapse, which starts by grouping reads that share a common set of splice junctions to cluster reads into putative isoforms (Tang et al., 2018; Brooks, 2019). Transcription start sites (TSS) and Transcription end sites (TES) are then determined by searching in 20 bp windows of putative TSS / TES and selecting the position with the highest number of 5′ / 3′ ends

that fall in that window (Tang et al., 2018; Brooks, 2019). Once a putative set of first pass isoforms is determined through these first two collapsing steps the sequence of these putative isoforms are extracted and reads are realigned to these sequences to ensure that there is enough read support for each isoform being reported (Tang et al., 2018).

Following this there are an additional two submodules: flair-quantify and flair-diffExp. These submodules, respectively, assign reads to isoforms for the purposes of quantifying the level of expression of each isoform, and then perform differential expression analysis on the isoforms to determine if isoforms are differentially expressed between conditions (Tang et al., 2018; Brooks, 2019). This is done by aligning reads to the reported isoforms from flair-collapse with minimap2 to determine isoform expression levels, followed by determining if isoforms are differentially expressed using the R package DESeq2 (Tang et al., 2018; Brooks, 2019).

1.5.5 TALON

Talon is the long read analysis pipeline utilized by the Encyclopedia of DNA Elements (ENCODE) project (Wyman et al., 2019). Like other long read analysis pipelines, TALON involves aligning reads to the genome and then includes an error correcting preprocessing step (Wyman et al., 2019). In the case of TALON this error correction step is encapsulated in a separate program called TranscriptClean (Wyman et al., 2019; Wyman and Mortazavi, 2019). TranscriptClean operates by removing and correcting indels smaller than some size threshold (which defaults to 5 bp), and replacing mismatches in the transcripts

with the reference genome base in that position (Wyman and Mortazavi, 2019). TranscriptClean also has a variant aware mode, which will only change indels and mismatches not present in a VCF file provided to the program, to allow for variant sequences to be maintained in the read (Wyman and Mortazavi, 2019). TranscriptClean also includes an optional non-canonical splice junction correction step, in which non-canonical splice junctions (NCSJs) are identified by checking the intron motifs of transcript splice sites (Wyman and Mortazavi, 2019). These splice sites are then compared to high confidence splice junctions provided by the user in the form of short read RNA-seq from the same sample or from a reference annotation and changed to match the known splice junction when the distance between the NCSJ and the nearest known junction is less than some fixed threshold which the TALON paper states represents the size of microindels in PacBio and ONT sequencing (Wyman and Mortazavi, 2019). Once preprocessing has been performed, the reads are then fed into the TALON pipeline proper (Wyman et al., 2019). This pipeline first establishes a database based on GENCODE data (Wyman et al., 2019). Alignments are then processed and compared to this database (Wyman et al., 2019). When novel transcript models are identified, they are added to the database and used in the comparisons with subsequent reads (Wyman et al., 2019). Each transcript model is assigned a transcript novelty type, which designates it as novel or previously known, and indicates how novel models differ from previously annotated models. These novelty types include: 1) Known transcript models, 2) ISMs or incomplete splice matches, which represent transcript models that partially match existing annotations but do not contain all the splice junctions

of an existing model 3) NIC – Novel in catalog, which represent novel transcript models that are combinations of previously existing splice junctions present in the existing annotation, 4) NNIC – Novel not in catalog, which represent novel transcript models that contain a splice junction not previously annotated to exist 5) Genomic transcript models which overlap an existing gene but do not share any of its splice donors or acceptors, 6) Antisense - which represent reads that overlap an existing gene but are oriented in the opposite direction and 7) Intergenic - which do not overlap with any known gene (Wyman et al., 2019). Once transcript models have been assigned to all reads, expression levels can be examined at the gene and transcript model level, novel transcript models can be identified, and transcript models can be visualized.

1.5.6 TALC - Transcription Aware Long-read correction

TALC is a hybrid long-read correction algorithm designed with long read RNA-seq in mind (Broseus et al., 2020). In TALC short, more accurate RNA-seq reads are used to construct weighted De-Bruijn graphs. Long reads are then compared to these De-Bruijn graphs, and stretches of common k-mers are used as anchor points for sections of each long read in the De-Bruijn graph. These common k-mers are assumed to be error free and are denoted "solid regions", while regions between common k-mers (denoted "weak regions") are corrected by exporing "consistent paths" from one anchor in the De-Bruijn graph to another. This correction algorithm operates under the assumption that k-mers that span a transcript should have consistent coverage across that transcript, except where multiple isoforms may alter coverage. Therefore,

"consistent path" extensions are defined by having kmer coverage consistent with the kmer coverage of the previous extension. This extension process identifies "split coverage events", where the consistent path is split into two or more competing paths when the coverage changes suddenly. These paths are then extended independently until the number of paths being considered rises above some arbitrary threshold, at which point all paths are compared to the long read being corrected, and those with the greatest edit distance to the long read are no longer considered. Paths that connect "solid regions" are then scored by edit distance to the long read, and the path most consistent with the long read is used to correct the "weak region". This results in long reads with errors corrected. It is worth noting, however, that this does not result in a transcriptome, and additional analysis is necessary to generate a transcriptome annotation.

1.5.7 RATTLE

RATTLE is a long-read *de novo* transcriptome assembler that operates independent of a reference genome (de la Rubia et al., 2020). RATTLE operates by first clustering long reads into gene level clusters, and then splitting these gene level clusters into isoform level clusters. A multiple sequence alignment of reads from each isoform is then generated using partial order alignment, and a consensus is generated from this multiple sequence alignment. These consensuses are then polished, and reported as the output transcriptome.

In this approach, gene level clusters are determined using a two step clustering algorithm. In the first step, reads sequence k-mers are extracted from each read

and used to generate a bitvector of length 4^k, where each position in the bit-vector represents a kmer, and each position corresponding to a k-mer present in the read is set to 1. For each initial comparison performed the bit-vectors are compared using the AND operator, resulting in the number of common k-mers. A similarity score is calculated and reads above some similarity score are then compared in a more computationally stringent manner. This second clustering step involves generating a list of kmers present in both reads sorted by thier position in the first read. The longest increasing subsequence for the positions in the second read is determined and used to calculate a second similarity metric. If reads are above some arbitrary threshold of the second similarity metric they are said to be in the same cluster, and thus originating from the same gene.

Rather than performing an all vs all comparison for every read, a process similar to canopy clustering is performed, in which an individual read is initialized as a cluster, and compared to every other read. All reads above the similarity threshold to this initial read are declared as part of the same cluster, and removed from consideration for future clusters. This process is repeated until every read is assigned to a cluster.

The similarity threshold is then gradually decreased, and clusters are merged based on the correspondence of a representative read from each cluster.

Once gene level clusters have been determined, they are split into isoform level clusters based on the relative distance between co-linear k-mers determined in the longest increasing subsequence calculations. Reads are split in to different transcripts if the variance of these distances between co-linear k-mers is greater

than some threshold set by the user.

Each isoform cluster is then used to generate a multiple sequence alignment using a SIMD partial order alignment tool. The consensus is calculated from this multiple sequence alignment using the base qualities from the FASTQ file input.

Consensuses are then polished by performing another 2-step greedy clustering on transcript clusters to define the final set of transcripts. From these final clusters consensuses are calculated and the isoform sequence, as well as the abundance of each isoform is reported in FASTA/Q format.

1.5.8 Trinity

Trinity is a short read de novo transcriptome assembler that operates independent of a reference genome (Grabherr et al., 2011). Trinity is so named because it operates in three major steps. First short reads are used to assemble a read data set, greedily searching through a k-mer graph to generate a set of linear contigs. These contigs are then pooled if they share at least 1 (k-1)-mer in common and used to construct De-Bruijn graphs for each pool. These De-Bruijn graphs are then trimmed of spurious edges, and resolved using reads to generate one sequence for each splice isoform. This approach is limited by the repetitive nature of the transcriptome as short reads cannot easily resolve cycles in De-Bruijn graphs.

1.5.9 Short read 3'UTR annotation approaches

Though much of the tools outlined above focus on obtaining a high quality annotation of the underlying structure of RNA transcripts and are focused primarily on CDS, the structure and sequence of the 3' untranslated regions (3'UTRs) of a gene are also extremely important for decoding the regulation of that gene. Determining 3'UTR structure and the cleavage and polyadenylation sites of RNA transcripts is a difficult process to address computationally for two main reasons 1) individual genes can contain several alternative cleavage and polyadenylation sites, 2) the available evidence indicates that cleavage does not reliably occur at the same nucleotide in every transcript, instead cleavage sites are broadly distributed in clusters around a given site (Hajarnavis et al., 2004; Mayr and Bartel, 2009; West et al., 2018). In addition, once sites have been identified, the short read lengths used by these studies means that associating a site with a gene often requires finding the nearest upstream gene and assigning the cleavage site to that gene (Mayr and Bartel, 2009; Mangone et al., 2010; Jan et al., 2011; Blazie et al., 2015, 2017).

Several computational approaches have been proposed and implemented in analysis of the *C. elegans* 3'UTR-ome in an attempt to utilize short read approaches to annotate 3'UTR structure.

In 3P-seq of the *C. elegans* transcriptome the reverse complement of sequencing reads were treated as candidate 3' tags and were first mapped to the genome allowing for the possibility of non-templated nucleotides on their 3' end (Jan et al., 2011). Sequences were kept if they aligned to only a single position in the genome and contained at least 2 3' adenosines, at least one of which

had to be non-templated (Jan et al., 2011). The 3′ most non-adenosine base of each remaining tag was then selected as a putative cleavage site (Jan et al., 2011). The genome was then cordoned off into genomic loci for the purposes of assigning tags to genes (Jan et al., 2011). In this cordoning scheme a gene's loci was defined as the region stretching from the 5′ most annotated base of that gene to the 5′ most annotated base of the next downstream gene on the same genomic strand (Jan et al., 2011). At each genomic loci candidate 3′ tags at each position in the loci were counted and sorted from most tags aligned to a given position to the fewest (Jan et al., 2011). Tags within 21 nt of the most abundant position were grouped into a cluster centered on that position and removed from the sorted list (Jan et al., 2011). This was repeated until all 3′ tags were assigned to a cluster belonging to a gene (Jan et al., 2011). The clusters were then evaluated using RNA-seq data to ensure that the maximum per base coverage of the putative 3′ UTR was less than 5 times the max of the upstream CDS, and that the median per base coverage was more than 0.05 the median per base coverage of the upstream CDS (Jan et al., 2011). The putative poly(A) addition sites that passed these filtering steps were then reported.

In Mangone et al. an alternative 3′UTR annotation approach was utilized. Mangone et al integrated several alternative sequencing datasets for profiling the 3′UTR-ome, each with unique pre-processing steps, the extracted 3′ positions of which were then pooled, clustered, and used to generate 3′UTR annotations (Mangone et al., 2010). Clustering was performed using an iterative local clustering procedure of their chromosomal coordinates. Putative

3'UTR sites that could be assigned to a gene were clustered on a gene by gene basis. The local maxima of each cluster was identified and then used as the representative poly(A) addition site for that cluster.

1.5.10 Short read transcription start site determination in *C. elegans*

Determining the transcription start site of genes is a critical step in annotating the transcriptome, made difficult in *C. elegans* by high rates of trans splicing. Alternative transcription start sites can lead to dramatic changes to the structure of an RNA transcript, and the resulting protein. Saito et al. and Chen et al. published two methods that attempted to overcome the trans splicing problem and define the TSS sites of *C. elegans*; the molecular biology involved in these techniques was explained above in Section 1.4.3.3 (Saito et al., 2013; Chen et al., 2013).

The computational analysis performed by Saito et al to determine TSS locations in the genome was quite straightforward. Upon mapping their short read sequencing data they started by computing the number of reads aligning each position to identify candidate TSS peaks (Saito et al., 2013). They then fit Gaussian Mixture Models to these candidate peaks, using an X-means clustering approach to fit these mixture models and determine the number of gaussian peaks to model based on a Bayesian Information Criterion, a concept from machine learning designed to optimize the number of parameters to include in a model such that the data is explained but the model is not overfit. Upon fitting their data, they trimmed peaks with fewer than 5 sequence tags supporting the peak, used the gaussian fit to estimate expression levels per

peak, and removed peaks that fell within 10 bp of an annotated SL1 trans splice site. Finally, they selected as representative TSS the peaks that were the highest expressed and fell within 3kb upstream of what they call 'gene start sites', which are not formally defined in the paper but are presumably the first base in each annotated gene.

The computational analysis in Chen et al. was focused on identifying Transcription Initiation Clusters or TICs and assigning these clusters to annotated TSS to determine the gene corresponding to each cluster (Chen et al., 2013). To do this, reads were mapped to the genome, and the strand specific alignment was used and the 5' ends of alignment were extracted. These 5' ends were combined together if they fell in the same position on the same strand, followed by combining these miniclusters together if they contained 2 or more reads by a single linkage approach in which miniclusters were combined if they fell within 50 nt of one another. Singleton 5' ends were added to clusters if they fell within the clustered region $\pm$25 nt. Clusters overlapping various classes of ncRNA (miRNA, rRNA, tRNA, snRNA, snoRNA and snlRNAs) were excluded from consideration, and the mode of each cluster was selected to define the putative TSS (ties were broken by selecting the mode closer to the median of the cluster). TSS were then assigned to genes in a multistep strand specific approach, which relied on scanning for an overlapping or nearest gene due to the short read lengths intrinsic to the sequencing used.

1.5.11 *trans*-splicing characterization computational approaches in *C. elegans*

Determining which genes are *trans* spliced in *C. elegans* and which genes are not using sequencing and computational approaches requires one to identify sequencing reads containing splice leader sequences and determine where these reads map in the genome. At least two approaches to solving this problem have been performed by different groups. Hillier et al. and later Allen et al. opted to create a 'trans-splice site database', in which all potential trans-splice sites (as predicted or annotated by Genefinder, TwinScan, or WormBase) were spliced *in silico* to SL1 and SL2 sequences (Hillier et al., 2009; Allen et al., 2011). RNAseq reads were then matched to the resulting database using the algorithm cross_match, and reads were assigned to a splice leader and location if the match score was better than all other match scores by two or more and better than all genome alignment scores by five or more. In this approach, trans-splice sites were only counted if at least one of the assigned reads had 9 or more bases matching the SL sequence.

In contrast, Tourasse et al., a meta analysis study working with many RNAseq datasets utilized an approach in which only reads that did not align to the genome in a first pass alignment were considered as trans-splicing candidates (Tourasse et al., 2017). These reads were then run through the program cutadapt to identify reads containing SL sequences and trim the reads of these sequences (Martin, 2011). The remaining portions of the reads were then mapped to the genome using TopHat2 to identify genomic trans splice sites (Kim et al., 2013). The reads that did not align to the genome were aligned

to the transcriptome by Bowtie2, identifying more trans splice sites (Langmead and Salzberg, 2012). Finally, the reads that were unmapped after this step were aligned to the genome using GSNAP, which identified additional genomic *trans* splice sites (Wu and Nacu, 2010). Once the mapping SL reads were identified, the relative position of *trans* splicing within the genome was determined as the 5′ most base before the SL sequence. These *trans* splice sites were then assigned to their overlapping gene, or in the case of intergenic *trans* splice sites, were assigned to the nearest downstream gene.

1.6 References

Akay A, Jordan D, Navarro IC, Wrzesinski T, Ponting CP, Miska EA, and Haerty W. 2019. Identification of functional long non-coding RNAs in c. elegans. *BMC Biol.* **17**: 14.

Albritton SE, Kranz AL, Rao P, Kramer M, Dieterich C, and Ercan S. 2014. Sex-biased gene expression and evolution of the x chromosome in nematodes. *Genetics* **197**: 865–883.

Allen MA, Hillier LW, Waterston RH, and Blumenthal T. 2011. A global analysis of c. elegans trans-splicing. *Genome Res.* **21**: 255–264.

Alton ZF and Hall DH. 2009. Introduction to c. elegans anatomy. https://www.wormatlas.org/hermaphrodite/introduction/mainframe.htm. Accessed: NA-NA-NA.

Ambros V. 2011. MicroRNAs and developmental timing. *Curr. Opin. Genet. Dev.* **21**: 511–517.

Ambros V, Lee RC, Lavanway A, Williams PT, and Jewell D. 2003. MicroRNAs and other tiny endogenous RNAs in c. elegans. *Curr. Biol.* **13**: 807–818.

Amrane S, Rebora K, Zniber I, Dupuy D, and Mackereth CD. 2014. Backbone-independent nucleic acid binding by splicing factor SUP-12 reveals key aspects of molecular recognition. *Nat. Commun.* **5**: 4595.

Andreassi C and Riccio A. 2009. To localize or not to localize: mRNA fate is in 3′ UTR ends. *Trends Cell Biol.* **19**: 465–474.

Ashe A, Sapetschnig A, Weick EM, Mitchell J, Bagijn MP, Cording AC, Doebley AL, Goldstein LD, Lehrbach NJ, Le Pen J, et al.. 2012. piRNAs can trigger a multigenerational epigenetic memory in the germline of c. elegans. *Cell* **150**: 88–99.

Atwater JA, Wisdom R, and Verma IM. 1990. Regulated mRNA stability. *Annu. Rev. Genet.* **24**: 519–541.

Bagijn MP, Goldstein LD, Sapetschnig A, Weick EM, Bouasker S, Lehrbach NJ, Simard MJ, and Miska EA. 2012. Function, targets, and evolution of caenorhabditis elegans piRNAs. *Science* **337**: 574–578.

Baralle FE and Giudice J. 2017. Alternative splicing as a regulator of development and tissue identity. *Nat. Rev. Mol. Cell Biol.* **18**: 437–451.

Barberan-Soler S, Lambert NJ, and Zahler AM. 2009. Global analysis of alternative splicing uncovers developmental regulation of nonsense-mediated decay in c. elegans. *RNA* **15**: 1652–1660.

Bartel DP. 2009. MicroRNAs: target recognition and regulatory functions. *Cell* **136**: 215–233.

Batista PJ, Ruby JG, Claycomb JM, Chiang R, Fahlgren N, Kasschau KD, Chaves DA, Gu W, Vasale JJ, Duan S, et al.. 2008. PRG-1 and 21U-RNAs interact to form the piRNA complex required for fertility in c. elegans. *Mol. Cell* **31**: 67–78.

Baugh LR. 2013. To grow or not to grow: nutritional control of development during caenorhabditis elegans L1 arrest. *Genetics* **194**: 539–555.

Bernstein E, Caudy AA, Hammond SM, and Hannon GJ. 2001. Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature* **409**: 363–366.

Billi AC, Fischer SEJ, and Kim JK. 2014. Endogenous RNAi pathways in c. elegans. *WormBook* pp. 1–49.

Billi AC, Freeberg MA, Day AM, Chun SY, Khivansara V, and Kim JK. 2013. A conserved upstream motif orchestrates autonomous, germline-enriched expression of caenorhabditis elegans piRNAs. *PLoS Genet.* **9**: e1003392.

Blazie SM, Babb C, Wilky H, Rawls A, Park JG, and Mangone M. 2015. Comparative RNA-Seq analysis reveals pervasive tissue-specific alternative polyadenylation in caenorhabditis elegans intestine and muscles. *BMC Biol.* **13**: 4.

Blazie SM, Geissel HC, Wilky H, Joshi R, Newbern J, and Mangone M. 2017. Alternative polyadenylation directs Tissue-Specific miRNA targeting in caenorhabditis elegans somatic tissues. *Genetics* **206**: 757–774.

Bleidorn C. 2016. Third generation sequencing: technology and its potential impact on evolutionary biodiversity research. *System. Biodivers.* **14**: 1–8.

Blumenthal T. 2005. Trans-splicing and operons. *WormBook* pp. 1–9.

Blumenthal T, Evans D, Link CD, Guffanti A, Lawson D, Thierry-Mieg J, Thierry-Mieg D, Chiu WL, Duke K, Kiraly M, et al.. 2002. A global analysis of caenorhabditis elegans operons. *Nature* **417**: 851–854.

Blumenthal T and Thomas J. 1988. Cis and trans mRNA splicing in c. elegans. *Trends Genet.* **4**: 305–308.

Boeck ME, Huynh C, Gevirtzman L, Thompson OA, Wang G, Kasper DM, Reinke V, Hillier LW, and Waterston RH. 2016. The time-resolved transcriptome of c. elegans. *Genome Res.* **26**: 1441–1450.

Borsani G, Tonlorenzi R, Simmler MC, Dandolo L, Arnaud D, Capra V, Grompe M, Pizzuti A, Muzny D, Lawrence C, et al.. 1991. Characterization of a murine gene expressed from the inactive X chromosome. *Nature* **351**: 325–329.

Bracht J, Hunter S, Eachus R, Weeks P, and Pasquinelli AE. 2004. Trans-splicing and polyadenylation of let-7 microRNA primary transcripts. *RNA* **10**: 1586–1594.

Brockdorff N, Ashworth A, Kay GF, McCabe VM, Norris DP, Cooper PJ, Swift S, and Rastan S. 1992. The product of the mouse xist gene is a 15 kb inactive x-specific transcript containing no conserved ORF and located in the nucleus. *Cell* **71**: 515–526.

Brooks A. 2019. flair.

Broseus L, Thomas A, Oldfield AJ, Severac D, Dubois E, and Ritchie W. 2020. TALC: Transcription aware long read correction.

Brow DA. 2002. Allosteric cascade of spliceosome activation. *Annu. Rev. Genet.* **36**: 333–360.

Brown C. 2019. London calling: Clive brown and team plenary. London Calling 2019.

Brown CJ, Hendrich BD, Rupert JL, Lafrenière RG, Xing Y, Lawrence J, and Willard HF. 1992. The human XIST gene: analysis of a 17 kb inactive x-specific RNA that contains conserved repeats and is highly localized within the nucleus. *Cell* **71**: 527–542.

Buermans HPJ and den Dunnen JT. 2014. Next generation sequencing technology: Advances and applications. *Biochim. Biophys. Acta* **1842**: 1932–1941.

Bussotti G, Leonardi T, Clark MB, Mercer TR, Crawford J, Malquori L, Notredame C, Dinger ME, Mattick JS, and Enright AJ. 2016. Improved definition of the mouse transcriptome via targeted RNA sequencing. *Genome Res.* **26**: 705–716.

Cáceres JF, Stamm S, Helfman DM, and Krainer AR. 1994. Regulation of alternative splicing in vivo by overexpression of antagonistic splicing factors. *Science* **265**: 1706–1709.

Cai Y, Yu X, Hu S, and Yu J. 2009. A brief review on the mechanisms of miRNA regulation. *Genomics Proteomics Bioinformatics* **7**: 147–154.

Calarco JA, Zhen M, and Blencowe BJ. 2011. Networking in a global world: establishing functional connections between neural splicing regulators and their target transcripts. *RNA* **17**: 775–791.

Cerritelli SM and Crouch RJ. 2009. Ribonuclease h: the enzymes in eukaryotes. *FEBS J.* **276**: 1494–1505.

Chang H, Lim J, Ha M, and Kim VN. 2014. TAIL-seq: genome-wide determination of poly(a) tail length and 3′ end modifications. *Mol. Cell* **53**: 1044–1052.

Chen RAJ, R A -, Down TA, Stempor P, Chen QB, Egelhofer TA, Hillier LW, Jeffers TE, and Ahringer J. 2013. The landscape of RNA polymerase II transcription initiation in c. elegans reveals promoter and enhancer architectures.

Cheng SC and Abelson J. 1987. Spliceosome assembly in yeast. *Genes Dev.* **1**: 1014–1027.

Conine CC, Batista PJ, Gu W, Claycomb JM, Chaves DA, Shirayama M, and Mello CC. 2010. Argonautes ALG-3 and ALG-4 are required for spermatogenesis-specific 26G-RNAs and thermotolerant sperm in caenorhabditis elegans. *Proc. Natl. Acad. Sci. U. S. A.* **107**: 3588–3593.

Cormen TH, Leiserson CE, Rivest RL, and Stein C. 2009. *Introduction to Algorithms*. The MIT Press, 3rd edition.

Cowling VH. 2009. Regulation of mRNA cap methylation. *Biochem. J* **425**: 295–302.

Cox GN and Hirsh D. 1985. Stage-specific patterns of collagen gene expression during development of caenorhabditis elegans. *Mol. Cell. Biol.* **5**: 363–372.

Craig NL, Cohen-Fix O, and Storz G. 2014. *Molecular Biology: Principles of Genome Function*. Oxford University Press.

Czech B, Munafò M, Ciabrelli F, Eastwood EL, Fabry MH, Kneuss E, and Hannon GJ. 2018. piRNA-Guided genome defense: From biogenesis to silencing. *Annu. Rev. Genet.* **52**: 131–157.

Dallaire A, Frédérick PM, and Simard MJ. 2018. Somatic and germline MicroRNAs form distinct silencing complexes to regulate their target mRNAs differently. *Dev. Cell* **47**: 239–247.e4.

Dalley BK and Golomb M. 1992. Gene expression in the caenorhabditis elegans dauer larva: developmental regulation of hsp90 and other genes. *Dev. Biol.* **151**: 80–90.

Das PP, Bagijn MP, Goldstein LD, Woolford JR, Lehrbach NJ, Sapetschnig A, Buhecha HR, Gilchrist MJ, Howe KL, Stark R, et al.. 2008. Piwi and piRNAs act upstream of an endogenous siRNA pathway to suppress tc3 transposon mobility in the caenorhabditis elegans germline. *Mol. Cell* **31**: 79–90.

Diag A, Schilling M, Klironomos F, Ayoub S, and Rajewsky N. 2018. Spatiotemporal m(i)RNA architecture and 3' UTR regulation in the c. elegans germline.

Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, et al.. 2012. Landscape of transcription in human cells. *Nature* **489**: 101–108.

Dupuy D, Li QR, Deplancke B, Boxem M, Hao T, Lamesch P, Sequerra R, Bosak S, Doucette-Stamm L, Hope IA, et al.. 2004. A first version of the caenorhabditis elegans promoterome. *Genome Res.* **14**: 2169–2175.

Eckmann CR, Rammelt C, and Wahle E. 2011. Control of poly(a) tail length. *Wiley Interdiscip. Rev. RNA* **2**: 348–361.

Eulalio A, Huntzinger E, and Izaurralde E. 2008. Getting to the root of miRNA-mediated gene silencing. *Cell* **132**: 9–14.

Fabian MR and Sonenberg N. 2012. The mechanics of miRNA-mediated gene silencing: a look under the hood of miRISC. *Nat. Struct. Mol. Biol.* **19**: 586–593.

Fischer SEJ, Montgomery TA, Zhang C, Fahlgren N, Breen PC, Hwang A, Sullivan CM, Carrington JC, and Ruvkun G. 2011. The ERI-6/7 helicase acts at the first stage of an siRNA amplification pathway that targets recent gene duplications. *PLoS Genet.* **7**: e1002369.

Fitzgerald M and Shenk T. 1981. The sequence 5′-AAUAAA-3′ forms part of the recognition site for polyadenylation of late SV40 mRNAs. *Cell* **24**: 251–260.

Furuichi Y, LaFiandra A, and Shatkin AJ. 1977. 5′-terminal structure and mRNA stability. *Nature* **266**: 235–239.

Gent JI, Lamm AT, Pavelec DM, Maniar JM, Parameswaran P, Tao L, Kennedy S, and Fire AZ. 2010. Distinct phases of siRNA synthesis in an endogenous RNAi pathway in c. elegans soma. *Mol. Cell* **37**: 679–689.

Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, Liu T, Yip KY, Robilotto R, Rechtsteiner A, Ikegami K, et al.. 2010. Integrative analysis of

the caenorhabditis elegans genome by the modENCODE project. *Science* **330**: 1775–1787.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, et al.. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**: 644–652.

Gracida X, Norris AD, and Calarco JA. 2016. Regulation of Tissue-Specific alternative splicing: C. elegans as a model system. In *RNA Processing: Disease and Genome-wide Probing* (ed. GW Yeo), pp. 229–261. Springer International Publishing, Cham.

Grishok A, Pasquinelli AE, Conte D, Li N, Parrish S, Ha I, Baillie DL, Fire A, Ruvkun G, and Mello CC. 2001. Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control c. elegans developmental timing. *Cell* **106**: 23–34.

Gu W, Lee HC, Chaves D, Youngman EM, Pazour GJ, Conte Jr D, and Mello CC. 2012. CapSeq and CIP-TAP identify pol II start sites and reveal capped small RNAs as c. elegans piRNA precursors. *Cell* **151**: 1488–1500.

Gu W, Shirayama M, Conte Jr D, Vasale J, Batista PJ, Claycomb JM, Moresco JJ, Youngman EM, Keys J, Stoltz MJ, et al.. 2009. Distinct argonaute-mediated 22G-RNA pathways direct genome surveillance in the c. elegans germline. *Mol. Cell* **36**: 231–244.

von der Haar T, Gross JD, Wagner G, and McCarthy JEG. 2004. The mRNA

cap-binding protein eIF4E in post-transcriptional gene expression. *Nat. Struct. Mol. Biol.* **11**: 503–511.

Haberle V and Stark A. 2018. Eukaryotic core promoters and the functional basis of transcription initiation. *Nat. Rev. Mol. Cell Biol.* **19**: 621–637.

Hajarnavis A, Korf I, and Durbin R. 2004. A probabilistic model of 3′ end formation in caenorhabditis elegans. *Nucleic Acids Res.* **32**: 3392–3399.

Hall SE, Beverly M, Russ C, Nusbaum C, and Sengupta P. 2010. A cellular memory of developmental history generates phenotypic diversity in c. elegans. *Curr. Biol.* **20**: 149–155.

Hall SE, Chirn GW, Lau NC, and Sengupta P. 2013. RNAi pathways contribute to developmental history-dependent phenotypic plasticity in c. elegans. *RNA* **19**: 306–319.

Han T, Manoharan AP, Harkins TT, Bouffard P, Fitzpatrick C, Chu DS, Thierry-Mieg D, Thierry-Mieg J, and Kim JK. 2009. 26G endo-siRNAs regulate spermatogenic and zygotic gene expression in caenorhabditis elegans. *Proc. Natl. Acad. Sci. U. S. A.* **106**: 18674–18679.

Han Li C and Chen Y. 2015. Small and long Non-Coding RNAs: Novel targets in perspective cancer therapy. *Curr. Genomics* **16**: 319–326.

Hansen KD, Lareau LF, Blanchette M, Green RE, Meng Q, Rehwinkel J, Gallusser FL, Izaurralde E, Rio DC, Dudoit S, et al.. 2009. Genome-wide identification of alternative splice forms down-regulated by nonsense-mediated mRNA decay in drosophila. *PLoS Genet.* **5**: e1000525.

Harris TW, Antoshechkin I, Bieri T, Blasiar D, Chan J, Chen WJ, De La Cruz N, Davis P, Duesbury M, Fang R, et al.. 2010. WormBase: a comprehensive resource for nematode research. *Nucleic Acids Res.* **38**: D463–7.

Hashimoto SI, Suzuki Y, Kasai Y, Morohoshi K, Yamada T, Sese J, Morishita S, Sugano S, and Matsushima K. 2004. 5'-end SAGE for the analysis of transcriptional start sites. *Nat. Biotechnol.* **22**: 1146–1149.

Hausen P and Stein H. 1970. Ribonuclease h: an enzyme degrading the RNA moiety of DNA-RNA hybrids. *Eur. J. Biochem.* .

Hillier LW, Coulson A, Murray JI, Bao Z, Sulston JE, and Waterston RH. 2005. Genomics in c. elegans: so many genes, such a little worm. *Genome Res.* **15**: 1651–1660.

Hillier LW, Reinke V, Green P, Hirst M, Marra MA, and Waterston RH. 2009. Massively parallel sequencing of the polyadenylated transcriptome of c. elegans. *Genome Res.* **19**: 657–666.

Hodgkin J, Horvitz HR, and Brenner S. 1979. Nondisjunction mutants of the nematode CAENORHABDITIS ELEGANS. *Genetics* **91**: 67–94.

Hollins C, Zorio DAR, MacMorris M, and Blumenthal T. 2005. U2AF binding selects for the high conservation of the c. elegans 3' splice site. *RNA* **11**: 248–253.

Holoch D and Moazed D. 2015. RNA-mediated epigenetic regulation of gene expression. *Nat. Rev. Genet.* **16**: 71–84.

Hoogstrate SW, Volkers RJ, Sterken MG, Kammenga JE, and Snoek LB. 2014. Nematode endogenous small RNA pathways. *Worm* **3**: e28234.

Huelga SC, Vu AQ, Arnold JD, Liang TY, Liu PP, Yan BY, Donohue JP, Shiue L, Hoon S, Brenner S, et al.. 2012. Integrative genome-wide analysis reveals cooperative regulation of alternative splicing by hnRNP proteins. *Cell Rep.* **1**: 167–178.

Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, Darbandi SF, Knowles D, Li YI, Kosmicki JA, Arbelaez J, Cui W, Schwartz GB, et al.. 2019. Predicting splicing from primary sequence with deep learning. *Cell* **176**: 535–548.e24.

Jan CH, Friedman RC, Ruby JG, and Bartel DP. 2011. Formation, regulation and evolution of caenorhabditis elegans 3'UTRs. *Nature* **469**: 97–101.

Johnston RJ and Hobert O. 2003. A microRNA controlling left/right neuronal asymmetry in caenorhabditis elegans. *Nature* **426**: 845–849.

Jonas S and Izaurralde E. 2015. Towards a molecular understanding of microRNA-mediated gene silencing. *Nat. Rev. Genet.* **16**: 421–433.

Kahvejian A, Svitkin YV, Sukarieh R, M'Boutchou MN, and Sonenberg N. 2005. Mammalian poly(a)-binding protein is a eukaryotic translation initiation factor, which acts via multiple mechanisms. *Genes Dev.* **19**: 104–113.

Karp X. 2018. Working with dauer larvae. *WormBook* **2018**: 1–19.

Karp X, Hammell M, Ow MC, and Ambros V. 2011. Effect of life history on microRNA expression during c. elegans development. *RNA* **17**: 639–651.

Kasai Y, Hashimoto SI, Yamada T, Sese J, Sugano S, Matsushima K, and Morishita S. 2005. 5'SAGE: 5'-end serial analysis of gene expression database. *Nucleic Acids Res.* **33**: D550–2.

Kawano T, Fujita M, and Sakamoto H. 2000. Unique and redundant functions of SR proteins, a conserved family of splicing factors, in caenorhabditis elegans development. *Mech. Dev.* **95**: 67–76.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, and Salzberg SL. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**: R36.

Kim JK, Gabel HW, Kamath RS, Tewari M, Pasquinelli A, Rual JF, Kennedy S, Dybbs M, Bertin N, Kaplan JM, et al.. 2005. Functional genomic analysis of RNA interference in c. elegans. *Science* **308**: 1164–1167.

Kim KW, Tang NH, Andrusiak MG, Wu Z, Chisholm AD, and Jin Y. 2018a. A neuronal piRNA pathway inhibits axon regeneration in c. elegans. *Neuron* **97**: 511–519.e6.

Kim W, Underwood RS, Greenwald I, and Shaye DD. 2018b. OrthoList 2: A new comparative genomic analysis of human and caenorhabditis elegans genes. *Genetics* **210**: 445–461.

Konarska MM and Sharp PA. 1986. Electrophoretic separation of complexes involved in the splicing of precursors to mRNAs. *Cell* **46**: 845–855.

Kono N and Arakawa K. 2019. Nanopore sequencing: Review of potential applications in functional genomics. *Dev. Growth Differ.* **61**: 316–326.

Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, and Pertea M. 2019. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* **20**: 278.

Kuroyanagi H, Kobayashi T, Mitani S, and Hagiwara M. 2006. Transgenic alternative-splicing reporters reveal tissue-specific expression profiles and regulation mechanisms in vivo. *Nat. Methods* **3**: 909–915.

Kuroyanagi H, Ohno G, Mitani S, and Hagiwara M. 2007. The fox-1 family and SUP-12 coordinately regulate tissue-specific alternative splicing in vivo. *Mol. Cell. Biol.* **27**: 8612–8621.

Kuroyanagi H, Watanabe Y, and Hagiwara M. 2013. CELF family RNA-binding protein UNC-75 regulates two sets of mutually exclusive exons of the unc-32 gene in neuron-specific manners in caenorhabditis elegans. *PLoS Genet.* **9**: e1003337.

Kuwasako K, Takahashi M, Unzai S, Tsuda K, Yoshikawa S, He F, Kobayashi N, Güntert P, Shirouzu M, Ito T, et al.. 2014. RBFOX and SUP-12 sandwich a G base to cooperatively regulate tissue-specific splicing. *Nat. Struct. Mol. Biol.* **21**: 778–786.

Lamesch P, Milstein S, Hao T, Rosenberg J, Li N, Sequerra R, Bosak S, Doucette-Stamm L, Vandenhaute J, Hill DE, et al.. 2004. C. elegans ORFeome version 3.1: increasing the coverage of ORFeome resources with improved gene predictions. *Genome Res.* **14**: 2064–2069.

Langmead B and Salzberg SL. 2012. Fast gapped-read alignment with bowtie 2. *Nat. Methods* **9**: 357–359.

Lee RYN, Howe KL, Harris TW, Arnaboldi V, Cain S, Chan J, Chen WJ, Davis P, Gao S, Grove C, et al.. 2018. WormBase 2017: molting into a new stage. *Nucleic Acids Res.* **46**: D869–D874.

Lee Y, Ahn C, Han J, Choi H, Kim J, Yim J, Lee J, Provost P, Rådmark O, Kim S, et al.. 2003. The nuclear RNase III drosha initiates microRNA processing. *Nature* **425**: 415–419.

Lee Y, Kim M, Han J, Yeom KH, Lee S, Baek SH, and Kim VN. 2004. MicroRNA genes are transcribed by RNA polymerase II. *EMBO J.* **23**: 4051–4060.

Legnini I, Alles J, Karaiskos N, Ayoub S, and Rajewsky N. 2019. FLAM-seq: full-length mRNA sequencing reveals principles of poly(a) tail length control. *Nat. Methods* **16**: 879–886.

Levin JZ, Yassour M, Adiconis X, Nusbaum C, Thompson DA, Friedman N, Gnirke A, and Regev A. 2010. Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat. Methods* **7**: 709–715.

Li J and Liu C. 2019. Coding or noncoding, the converging concepts of RNAs. *Front. Genet.* **10**: 496.

Li Y and Kiledjian M. 2010. Regulation of mRNA decapping. *Wiley Interdiscip. Rev. RNA* **1**: 253–265.

Lim J, Lee M, Son A, Chang H, and Kim VN. 2016. mTAIL-seq reveals dynamic poly(a) tail regulation in oocyte-to-embryo development. *Genes Dev.* **30**: 1671–1682.

Lima SA, Chipman LB, Nicholson AL, Chen YH, Yee BA, Yeo GW, Coller J, and Pasquinelli AE. 2017. Short poly(a) tails are a conserved feature of highly expressed genes. *Nat. Struct. Mol. Biol.* **24**: 1057–1063.

Lipton ZC, Berkowitz J, and Elkan C. 2015. A critical review of recurrent neural networks for sequence learning .

Liu Z, Luyten I, Bottomley MJ, Messias AC, Houngninou-Molango S, Sprangers R, Zanier K, Krämer A, and Sattler M. 2001. Structural basis for recognition of the intron branch site RNA by splicing factor 1. *Science* **294**: 1098–1102.

Longman D, Johnstone IL, and Cáceres JF. 2000. Functional characterization of SR and SR-related genes in caenorhabditis elegans. *EMBO J.* **19**: 1625–1637.

Longman D, McGarvey T, McCracken S, Johnstone IL, Blencowe BJ, and Cáceres JF. 2001. Multiple interactions between SRm160 and SR family proteins in enhancer-dependent splicing and development of c. elegans. *Curr. Biol.* **11**: 1923–1933.

Lu H, Giordano F, and Ning Z. 2016. Oxford nanopore MinION sequencing and genome assembly. *Genomics Proteomics Bioinformatics* **14**: 265–279.

Lundquist EA and Herman RK. 1994. The mec-8 gene of caenorhabditis elegans affects muscle and sensory neuron function and interacts with three other genes: unc-52, smu-1 and smu-2. *Genetics* **138**: 83–101.

Lundquist EA, Herman RK, Rogalski TM, Mullen GP, Moerman DG, and Shaw JE. 1996. The mec-8 gene of c. elegans encodes a protein with two RNA

recognition motifs and regulates alternative splicing of unc-52 transcripts. *Development* **122**: 1601–1610.

Luo C, Tsementzi D, Kyrpides N, Read T, and Konstantinidis KT. 2012. Direct comparisons of illumina vs. roche 454 sequencing technologies on the same microbial community DNA sample. *PLoS One* **7**: e30087.

Luteijn MJ, van Bergeijk P, Kaaij LJT, Almeida MV, Roovers EF, Berezikov E, and Ketting RF. 2012. Extremely stable piwi-induced gene silencing in caenorhabditis elegans. *EMBO J.* **31**: 3422–3430.

MacMorris M, Kumar M, Lasda E, Larsen A, Kraemer B, and Blumenthal T. 2007. A novel family of c. elegans snRNPs contains proteins associated with trans-splicing. *RNA* **13**: 511–520.

Maitra RD, Kim J, and Dunbar WB. 2012. Recent advances in nanopore sequencing. *Electrophoresis* **33**: 3418–3428.

Malone CD and Hannon GJ. 2009. Small RNAs as guardians of the genome. *Cell* **136**: 656–668.

Mangone M, Manoharan AP, Thierry-Mieg D, Thierry-Mieg J, Han T, Mackowiak SD, Mis E, Zegar C, Gutwein MR, Khivansara V, et al.. 2010. The landscape of c. elegans 3'UTRs. *Science* **329**: 432–435.

Marchese FP, Raimondi I, and Huarte M. 2017. The multidimensional mechanisms of long noncoding RNA function. *Genome Biol.* **18**: 206.

Mardis ER. 2008. Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.* **9**: 387–402.

Mardis ER. 2013. Next-generation sequencing platforms. *Annu. Rev. Anal. Chem.* **6**: 287–303.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**: 10–12.

Mathé C, Sagot MF, Schiex T, and Rouzé P. 2002. Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Res.* **30**: 4103–4117.

Mayr C and Bartel DP. 2009. Widespread shortening of 3′ UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell* **138**: 673–684.

Mazroui R, Puoti A, and Krämer A. 1999. Splicing factor SF1 from drosophila and caenorhabditis: presence of an n-terminal RS domain and requirement for viability. *RNA* **5**: 1615–1631.

Mercer TR, Dinger ME, and Mattick JS. 2009. Long non-coding RNAs: insights into functions. *Nat. Rev. Genet.* **10**: 155–159.

Montgomery TA, Rim YS, Zhang C, Dowen RH, Phillips CM, Fischer SEJ, and Ruvkun G. 2012. PIWI associated siRNAs and piRNAs specifically require the caenorhabditis elegans HEN1 ortholog henn-1. *PLoS Genet.* **8**: e1002616.

Morris KV and Mattick JS. 2014. The rise of regulatory RNA. *Nat. Rev. Genet.* **15**: 423–437.

Morrison M, Harris KS, and Roth MB. 1997. smg mutants affect the expression of alternatively spliced SR protein mRNAs in caenorhabditis elegans. *Proc. Natl. Acad. Sci. U. S. A.* **94**: 9782–9785.

Motta-Mena LB, Heyd F, and Lynch KW. 2010. Context-dependent regulatory mechanism of the splicing factor hnRNP L. *Mol. Cell* **37**: 223–234.

Muir VS, Gasch AP, and Anderson P. 2018. The substrates of Nonsense-Mediated mRNA decay in caenorhabditis elegans. *G3* **8**: 195–205.

Murthy KG and Manley JL. 1995. The 160-kd subunit of human cleavage-polyadenylation specificity factor coordinates pre-mRNA 3′-end formation. *Genes Dev.* **9**: 2672–2683.

Nam JW and Bartel DP. 2012. Long noncoding RNAs in c. elegans. *Genome Res.* **22**: 2529–2540.

Ni JZ, Grate L, Donohue JP, Preston C, Nobida N, O'Brien G, Shiue L, Clark TA, Blume JE, and Ares Jr M. 2007. Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay. *Genes Dev.* **21**: 708–718.

Ni P, Huang N, Zhang Z, Wang DP, Liang F, Miao Y, Xiao CL, Luo F, and Wang J. 2019. DeepSignal: detecting DNA methylation state from nanopore sequencing reads using deep-learning. *Bioinformatics* **35**: 4586–4595.

Nudel U, Soreq H, and Littauer UZ. 1976. Globin mRNA species containing poly(a) segments of different lengths. their functional stability in xenopus oocytes. *Eur. J. Biochem.* **64**: 115–121.

Olsen PH and Ambros V. 1999. The lin-4 regulatory RNA controls developmental timing in caenorhabditis elegans by blocking LIN-14 protein synthesis after the initiation of translation. *Dev. Biol.* **216**: 671–680.

Oxford Nanopore Technologies. 2019. flappie.

Packer JS, Zhu Q, Huynh C, Sivaramakrishnan P, Preston E, Dueck H, Stefanik D, Tan K, Trapnell C, Kim J, et al.. 2019. A lineage-resolved molecular atlas of c. elegans embryogenesis at single-cell resolution. *Science* **365**.

Pan Q, Shai O, Lee LJ, Frey BJ, and Blencowe BJ. 2008. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* **40**: 1413–1415.

Paraskevopoulou MD and Hatzigeorgiou AG. 2016. Analyzing MiRNA-LncRNA interactions. *Methods Mol. Biol.* **1402**: 271–286.

Parkinson J and Blaxter M. 2009. Expressed sequence tags: an overview. *Methods Mol. Biol.* **533**: 1–12.

Paule MR and White RJ. 2000. Survey and summary: transcription by RNA polymerases I and III. *Nucleic Acids Res.* **28**: 1283–1298.

Pavelec DM, Lachowiec J, Duchaine TF, Smith HE, and Kennedy S. 2009. Requirement for the ERI/DICER complex in endogenous RNA interference and sperm development in caenorhabditis elegans. *Genetics* **183**: 1283–1295.

Payne A, Holmes N, Rakyan V, and Loose M. 2019. BulkVis: a graphical viewer for oxford nanopore bulk FAST5 files. *Bioinformatics* **35**: 2193–2198.

Peng JC and Lin H. 2013. Beyond transposons: the epigenetic and somatic functions of the Piwi-piRNA mechanism. *Curr. Opin. Cell Biol.* **25**: 190–194.

Pertea M, Kim D, Pertea GM, Leek JT, and Salzberg SL. 2016. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and ballgown. *Nat. Protoc.* **11**: 1650–1667.

Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, and Salzberg SL. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**: 290–295.

Pikielny CW, Rymond BC, and Rosbash M. 1986. Electrophoresis of ribonucleoproteins reveals an ordered assembly pathway of yeast splicing complexes. *Nature* **324**: 341–345.

Pillai RS, Bhattacharyya SN, Artus CG, Zoller T, Cougot N, Basyuk E, Bertrand E, and Filipowicz W. 2005. Inhibition of translational initiation by let-7 MicroRNA in human cells. *Science* **309**: 1573–1576.

Porrua O and Libri D. 2015. Transcription termination and the control of the transcriptome: why, where and how to stop. *Nat. Rev. Mol. Cell Biol.* **16**: 190–202.

Preiss T, Muckenthaler M, and Hentze MW. 1998. Poly(A)-tail-promoted translation in yeast: implications for translational control. *RNA* **4**: 1321–1331.

Proudfoot NJ. 2011. Ending the message: poly(a) signals then and now. *Genes Dev.* **25**: 1770–1782.

Quinn JJ and Chang HY. 2016. Unique features of long non-coding RNA biogenesis and function. *Nat. Rev. Genet.* **17**: 47–62.

Rabiner LR. 1989. A tutorial on hidden markov models and selected applications in speech recognition. *Proc. IEEE* **77**: 257–286.

Ramanathan A, Robb GB, and Chan SH. 2016. mRNA capping: biological functions and applications. *Nucleic Acids Res.* **44**: 7511–7526.

Rang FJ, Kloosterman WP, and de Ridder J. 2018. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome Biol.* **19**: 90.

Reboul J, Vaglio P, Tzellas N, Thierry-Mieg N, Moore T, Jackson C, Shin-i T, Kohara Y, Thierry-Mieg D, Thierry-Mieg J, et al.. 2001. Open-reading-frame sequence tags (OSTs) support the existence of at least 17,300 genes in c. elegans. *Nat. Genet.* **27**: 332–336.

Reinhart BJ, Slack FJ, Basson M, Pasquinelli AE, Bettinger JC, Rougvie AE, Horvitz HR, and Ruvkun G. 2000. The 21-nucleotide let-7 RNA regulates developmental timing in caenorhabditis elegans. *Nature* **403**: 901–906.

Rhoads A and Au KF. 2015. PacBio sequencing and its applications. *Genomics Proteomics Bioinformatics* **13**: 278–289.

Riddle DL, Blumenthal T, Meyer BJ, and Priess JR. 1997. *Cis -Splicing in Worms*. Cold Spring Harbor Laboratory Press.

Robert VJP, Sijen T, van Wolfswinkel J, and Plasterk RHA. 2005. Chromatin and RNAi factors protect the c. elegans germline against repetitive sequences. *Genes Dev.* **19**: 782–787.

Ross LH, Freedman JH, and Rubin CS. 1995. Structure and expression of novel spliced leader RNA genes in caenorhabditis elegans. *J. Biol. Chem.* **270**: 22066–22075.

Rouget C, Papin C, Boureux A, Meunier AC, Franco B, Robine N, Lai EC, Pelisson A, and Simonelig M. 2010. Maternal mRNA deadenylation and decay by the piRNA pathway in the early drosophila embryo. *Nature* **467**: 1128–1132.

de la Rubia I, Indi JA, Carbonell S, Lagarde J, Mar Albà M, and Eyras E. 2020. Reference-free reconstruction and quantification of transcriptomes from long-read sequencing.

Ruby JG, Jan C, Player C, Axtell MJ, Lee W, Nusbaum C, Ge H, and Bartel DP. 2006. Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in c. elegans. *Cell* **127**: 1193–1207.

Ruby SW. 1997. Dynamics of the U1 small nuclear ribonucleoprotein during yeast spliceosome assembly. *J. Biol. Chem.* **272**: 17333–17341.

Saito TL, Hashimoto SI, Gu SG, Morton JJ, Stadler M, Blumenthal T, Fire A, and Morishita S. 2013. The transcription start site landscape of c. elegans. *Genome Res.* **23**: 1348–1361.

Schwartz DC and Parker R. 2000. mRNA decapping in yeast requires dissociation of the cap binding protein, eukaryotic translation initiation factor 4E. *Mol. Cell. Biol.* **20**: 7933–7942.

Selenko P, Gregorovic G, Sprangers R, Stier G, Rhani Z, Krämer A, and Sattler M. 2003. Structural basis for the molecular recognition between human splicing factors U2AF65 and SF1/mBBP. *Mol. Cell* **11**: 965–976.

Seraphin B and Rosbash M. 1989. Identification of functional U1 snRNA-pre-mRNA complexes committed to spliceosome assembly and splicing. *Cell* **59**: 349–358.

Shaye DD and Greenwald I. 2011. OrthoList: a compendium of c. elegans genes with human orthologs. *PLoS One* **6**: e20085.

Shimotohno K, Kodama Y, Hashimoto J, and Miura KI. 1977. Importance of 5'-terminal blocking structure to stabilize mRNA in eukaryotic protein synthesis. *Proc. Natl. Acad. Sci. U. S. A.* **74**: 2734–2738.

Shirayama M, Seth M, Lee HC, Gu W, Ishidate T, Conte Jr D, and Mello CC. 2012. piRNAs initiate an epigenetic memory of nonself RNA in the c. elegans germline. *Cell* **150**: 65–77.

Simpson JT, Workman RE, Zuzarte PC, David M, Dursi LJ, and Timp W. 2017. Detecting DNA cytosine methylation using nanopore sequencing. *Nat. Methods* **14**: 407–410.

Smith CW and Valcárcel J. 2000. Alternative pre-mRNA splicing: the logic of combinatorial control. *Trends Biochem. Sci.* **25**: 381–388.

Spieth J, Genome Sequencing Center, Washington University School of Medicine, Louis S, and Usa MO. 2014. Overview of gene structure in c. elegans.

Srivastava A, George J, and Karuturi RKM. 2019. Transcriptome analysis. In *Encyclopedia of Bioinformatics and Computational Biology* (eds. S Ranganathan, M Gribskov, K Nakai, and C Schönbach), pp. 792–805. Academic Press, Oxford.

Stark R, Grzelak M, and Hadfield J. 2019. RNA sequencing: the teenage years. *Nat. Rev. Genet.* **20**: 631–656.

Stoiber M, Quick J, Egan R, Lee JE, Celniker S, Neely RK, Loman N, Pennacchio LA, and Brown J. 2017. De novo identification of DNA modifications enabled by Genome-Guided nanopore signal processing.

Subtelny AO, Eichhorn SW, Chen GR, Sive H, and Bartel DP. 2014. Poly(A)-tail profiling reveals an embryonic switch in translational control. *Nature* **508**: 66–71.

Szostak E and Gebauer F. 2013. Translational control by 3'-UTR-binding proteins. *Brief. Funct. Genomics* **12**: 58–65.

Takagi T, Walker AK, Sawa C, Diehn F, Takase Y, Blackwell TK, and Buratowski S. 2003. The *Caenorhabditis elegans* mRNA 5'-capping enzyme: IN VITRO AND IN VIVO CHARACTERIZATION. *J. Biol. Chem.* **278**: 14174–14184.

Tang AD, Soulette CM, van Baren MJ, Hart K, and others. 2018. Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *bioRxiv* .

Tourasse NJ, Millet JRM, and Dupuy D. 2017. Quantitative RNA-seq meta-analysis of alternative exon usage in c. elegans. *Genome Res.* **27**: 2120–2128.

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, and Pachter L. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and cufflinks. *Nat. Protoc.* **7**: 562–578.

Vasale JJ, Gu W, Thivierge C, Batista PJ, Claycomb JM, Youngman EM, Duchaine TF, Mello CC, and Conte Jr D. 2010. Sequential rounds of RNA-dependent RNA transcription drive endogenous small-RNA biogenesis in the ERGO-1/Argonaute pathway. *Proc. Natl. Acad. Sci. U. S. A.* **107**: 3582–3587.

Vella MC and Slack FJ. 2005. *C. elegans microRNAs*. WormBook.

Walhout AJM, Temple GF, Brasch MA, Hartley JL, Lorson MA, van den Heuvel S, and Vidal M. 2000. [34] GATEWAY recombinational cloning: Application to the cloning of large numbers of open reading frames or ORFeomes. In *Methods in Enzymology* (eds. J Thorner, SD Emr, and JN Abelson), volume 328, pp. 575–IN7. Academic Press.

Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, and Burge CB. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**: 470–476.

Wang J and Kim SK. 2003. Global analysis of dauer gene expression in caenorhabditis elegans. *Development* **130**: 1621–1634.

Wang Z and Burge CB. 2008. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA* **14**: 802–813.

Wani S and Kuroyanagi H. 2017. An emerging model organism caenorhabditis elegans for alternative pre-mRNA processing in vivo. *WIREs RNA* **8**: 860.

Weiser N. 2019. *Multigenerational Regulation of the C. elegans Chromatin Landscape by Germline Small RNAs*. Ph.D. thesis.

West SM, Mecenas D, Gutwein M, Aristizábal-Corrales D, Piano F, and Gunsalus KC. 2018. Developmental dynamics of gene expression and alternative polyadenylation in the caenorhabditis elegans germline. *Genome Biol.* **19**: 8.

Wick RR, Judd LM, and Holt KE. 2019. Performance of neural network basecalling tools for oxford nanopore sequencing. *Genome Biol.* **20**: 129.

Wilusz JE, Sunwoo H, and Spector DL. 2009. Long noncoding RNAs: functional surprises from the RNA world. *Genes Dev.* **23**: 1494–1504.

Wollerton MC, Gooding C, Wagner EJ, Garcia-Blanco MA, and Smith CWJ. 2004. Autoregulation of polypyrimidine tract binding protein by alternative splicing leading to nonsense-mediated decay. *Mol. Cell* **13**: 91–100.

Workman RE, Tang AD, Tang PS, Jain M, Tyson JR, Razaghi R, Zuzarte PC, Gilpatrick T, Payne A, Quick J, et al.. 2019. Nanopore native RNA sequencing of a human poly(a) transcriptome. *Nat. Methods* **16**: 1297–1305.

WormBase web site. 2018. http://wormbase.org, release WS265.

Wu TD and Nacu S. 2010. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**: 873–881.

Wyman D, Balderrama-Gutierrez G, Reese F, Jiang S, Rahmanian S, Zeng W, Williams B, Trout D, England W, Chu S, et al.. 2019. A technology-agnostic long-read analysis pipeline for transcriptome discovery and quantification.

Wyman D and Mortazavi A. 2019. TranscriptClean: variant-aware correction of indels, mismatches and splice junctions in long-read transcripts. *Bioinformatics* **35**: 340–342.

Yang YF, Zhang X, Ma X, Zhao T, Sun Q, Huan Q, Wu S, Du Z, and Qian W. 2017. Trans-splicing enhances translational efficiency in c. elegans. *Genome Res.* **27**: 1525–1535.

Yigit E, Batista PJ, Bei Y, Pang KM, Chen CCG, Tolia NH, Joshua-Tor L, Mitani S, Simard MJ, and Mello CC. 2006. Analysis of the c. elegans argonaute family reveals that distinct argonautes act sequentially during RNAi. *Cell* **127**: 747–757.

Zahler AM. 2005. *Alternative splicing in C. elegans*. WormBook.

Zahler AM. 2018. *Pre-mRNA splicing and its regulation in Caenorhabditis elegans*. WormBook.

Zhao HQ, Zhang P, Gao H, He X, Dou Y, Huang AY, Liu XM, Ye AY, Dong MQ, and Wei L. 2015. Profiling the RNA editomes of wild-type c. elegans and ADAR mutants. *Genome Res.* **25**: 66–75.

Zimin AV, Marçais G, Puiu D, Roberts M, Salzberg SL, and Yorke JA. 2013. The MaSuRCA genome assembler. *Bioinformatics* **29**: 2669–2677.

Zorio DA and Blumenthal T. 1999a. Both subunits of U2AF recognize the 3′ splice site in caenorhabditis elegans. *Nature* **402**: 835–838.

Zorio DA and Blumenthal T. 1999b. U2AF35 is encoded by an essential gene clustered in an operon with RRM/cyclophilin in caenorhabditis elegans. *RNA* **5**: 487–494.

Zorio DA, Cheng NN, Blumenthal T, and Spieth J. 1994. Operons as a common form of chromosomal organization in c. elegans. *Nature* **372**: 270–272.

Zorio DA, Lea K, and Blumenthal T. 1997. Cloning of caenorhabditis U2AF65: an alternatively spliced RNA containing a novel exon. *Mol. Cell. Biol.* **17**: 946–953.

Chapter 2

The full-length transcriptome of *C. elegans* using direct RNA sequencing

The following chapter has been published in the journal Genome Research under the title "The full-length transcriptome of *C. elegans* using direct RNA sequencing." It is reproduced here, in its entirety (including supplemental materials) in accordance with the Genome Research license to publish document, which states:

> "1. Ownership of copyright remains with the Authors (except US Government employees), and provided that, when reproducing the Article or extracts from it, the Authors acknowledge first publication in Genome Research, the Authors retain the following non-exclusive rights:
> a) To reproduce the Article in whole or in part in any printed volume (book or thesis) of which they are the author(s)."

And in accordance with the handbook for the Johns Hopkins Cellular Molecular Developmental Biology and Biophysics department which states:

> "It is customary to include first-author papers in a thesis as long as you include a description at the end of the chapter about your contributions vs. other author contributions."

2.1 Citation

Roach NP, Sadowski N, Alessi AF, Timp W, Taylor J, and Kim JK. 2020. The full-length transcriptome of c. elegans using direct RNA sequencing. *Genome Res.* **30**: 299–312.

2.2 Author Contributions

Amelia F. Alessi collected the developmental samples for *C. elegans* and isolated the RNA to be sequenced, and Norah Sadowski performed the sequencing. Nathan P. Roach performed all of the sequencing analysis and wrote the manuscript in consultation with James Taylor, John K. Kim, and Winston Timp. Winston Timp, James Taylor, and John K. Kim conceived of and supervised various aspects of the project. All authors reviewed the manuscript.

2.3 Introduction

The nematode *Caenorhabditis elegans* is an ideal experimental model organism due to its compact, well-annotated genome (The C. elegans Sequencing Consortium, 1998; Wilson, 1999; Hillier et al., 2005; Gerstein et al., 2010), invariant cell lineage (Sulston et al., 1983), and wide array of molecular methods. Our current understanding of the *C. elegans* transcriptome has been determined with EST- and cDNA-based libraries, and Illumina-based cDNA and RNA sequencing (Walhout et al., 2000; Reboul et al., 2001; Lamesch et al., 2004; Hillier et al., 2009; Gerstein et al., 2010; Spieth et al., 2014; Tourasse et al., 2017). Most coding sequences (CDSs) span more than 600 nt (excluding introns), and the typical *C. elegans* gene contains 6.4 coding exons on average (Spieth et al., 2014).

3′ untranslated regions (3′UTRs) are critically important features of mRNA transcripts that contain binding sites for RNA-binding proteins and small noncoding RNAs (Cai et al., 2009; Szostak and Gebauer, 2013). Regulation of 3′-UTR length can therefore have profound impact on mRNA expression, stability, and localization (Kuersten and Goodwin, 2003; Andreassi and Riccio, 2009; Mayr and Bartel, 2009). Large-scale sequencing of the *C. elegans* 3′UTRs revealed median lengths of 130–140 nt, with an average length of ~211 nt, although 3′-UTR length distributions have been shown to vary by cell and tissue type (e.g., oocytes have a median 3′-UTR length of ~157 nt) (Mangone et al., 2010; Jan et al., 2011; West et al., 2018). In addition, poly(A) tails in *C. elegans* have a median length of ~57 nt at the L4 stage, and short poly(A) tail lengths are a feature of highly expressed genes (Lima et al., 2017).

The average transcript in *C. elegans* is significantly longer than the maximum possible read length of Illumina sequencing. Therefore, current approaches to annotate the full-length structure of the average *C. elegans* transcript isoform rely on manual curation of gene models based on a variety of data types, while more generally computational approaches to assemble transcript structures from bulk, short-read sequencing data utilize computationally expensive and imperfect inference (Williams et al., 2011; Trapnell et al., 2012; Spieth et al., 2014; Pertea et al., 2015). Calculating poly(A) tail lengths requires a sequencing approach capable of resolving long homopolymers, and determining 3′-UTR structures requires an experimental or computational means of determining which reads reflect the 3′-most base included in the transcript before cleavage and polyadenylation. The specialized protocols and analyses used to measure poly(A) tail length and identify 3′UTRs with short-read sequencing approaches cannot directly link these measurements to their splice isoform of origin and, in the case of 3′-UTR identification, instead rely on assigning putative cleavage sites to the nearest overlapping or upstream gene (Mangone et al., 2010; Jan et al., 2011; Chang et al., 2014; Subtelny et al., 2014; Blazie et al., 2017; Diag et al., 2018).

Nanopore sequencing, in contrast, has no theoretical upper limit to read length and is capable of sequencing transcripts from end to end at a single molecule level (Garalde et al., 2018; Jenjaroenpun et al., 2018; Workman et al., 2019). Nanopore-based sequencing methods have been used to annotate transcriptome structure in a variety of organisms ranging from the relatively simple *Saccharomyces cerevisiae* to complex human cell lines (Byrne et al., 2017; Bayega

et al., 2018; Garalde et al., 2018; Jenjaroenpun et al., 2018; Tang et al., 2018; Volden et al., 2018; Kadobianskyi et al., 2019; Sessegolo et al., 2019; Workman et al., 2019). In nanopore based direct RNA sequencing (dRNA-seq), RNA reads are captured by the 3′ end of their poly(A) tail and sequenced in the 3′ to 5′ direction natively, thus directly measuring the RNA molecule. The full length of the poly(A) tail is sequenced and, using a trained hidden Markov model, the length of the poly(A) tail for each read can be estimated (Workman et al., 2019). The 3′-most base in the alignment should reflect the true cleavage and polyadenylation site for the full transcript represented by that read, provided that base-calling, trimming of poly(A) and adapter sequences, and alignment had sufficient precision. Despite these advantages, adoption of dRNA-seq and other nanopore-based sequencing methods is hindered due to the technology's high error rates and the relative lack of bioinformatics tools and analysis pipelines designed for long, error-rich reads.

In this study, we have generated an atlas of postembryonic transcript structure using dRNA-seq to sequence RNA extracted from the major stages of the *C. elegans* developmental life cycle. We provide full-length support for both previously annotated and novel transcript splice isoforms. Furthermore, we identify and characterize 3′UTRs and compare these to known 3′-UTR data sets. We also estimate poly(A) tail lengths for our reads and examine their length characteristics across development. Finally, we have made this data available both in raw formats and as a custom track hub.

2.4 Results

2.4.1 Collection and sequencing of developmentally staged *C. elegans*

To capture the diversity of transcript isoforms expressed across *C. elegans* development, we created dRNA-seq libraries in technical duplicates from larval stages L1 to L4, as well as young and mature hermaphrodite adults (Fig. 2.1A; Corsi et al. (2015)). Because wild-type *C. elegans* exists primarily as hermaphrodites with spontaneous males (<0.5%) emerging in the population through chromosome nondisjunction, we also obtained a male-enriched sample using a *him-8* mutant that disrupts X Chromosome segregation (Hodgkin et al., 1979; Broverman and Meneely, 1994; Phillips et al., 2005). We further enriched for the male subpopulation by filtering them through a 35-µm mesh that allows the males to be collected in the filtrate.

Libraries were generated from RNA isolated by TRI Reagent (Ambion), poly(A)-selected, and prepared for sequencing following the Oxford Nanopore Technologies SQK-RNA001 kit protocol with the exception of using SuperScript IV (Thermo Fisher Scientific) in the optional reverse transcription step. The libraries were sequenced on an Oxford Nanopore Technologies GridION X5 (model # GRD-X5B002). Base-calling and adapter trimming of the reads was performed using poreplex (running albacore) (`https://github.com/hyeshik/poreplex`), resulting in over 540,000 reads that passed base-calling quality control for each developmental stage sequenced, and 5.54 million total reads (Supplemental Table S2.1). Reads had

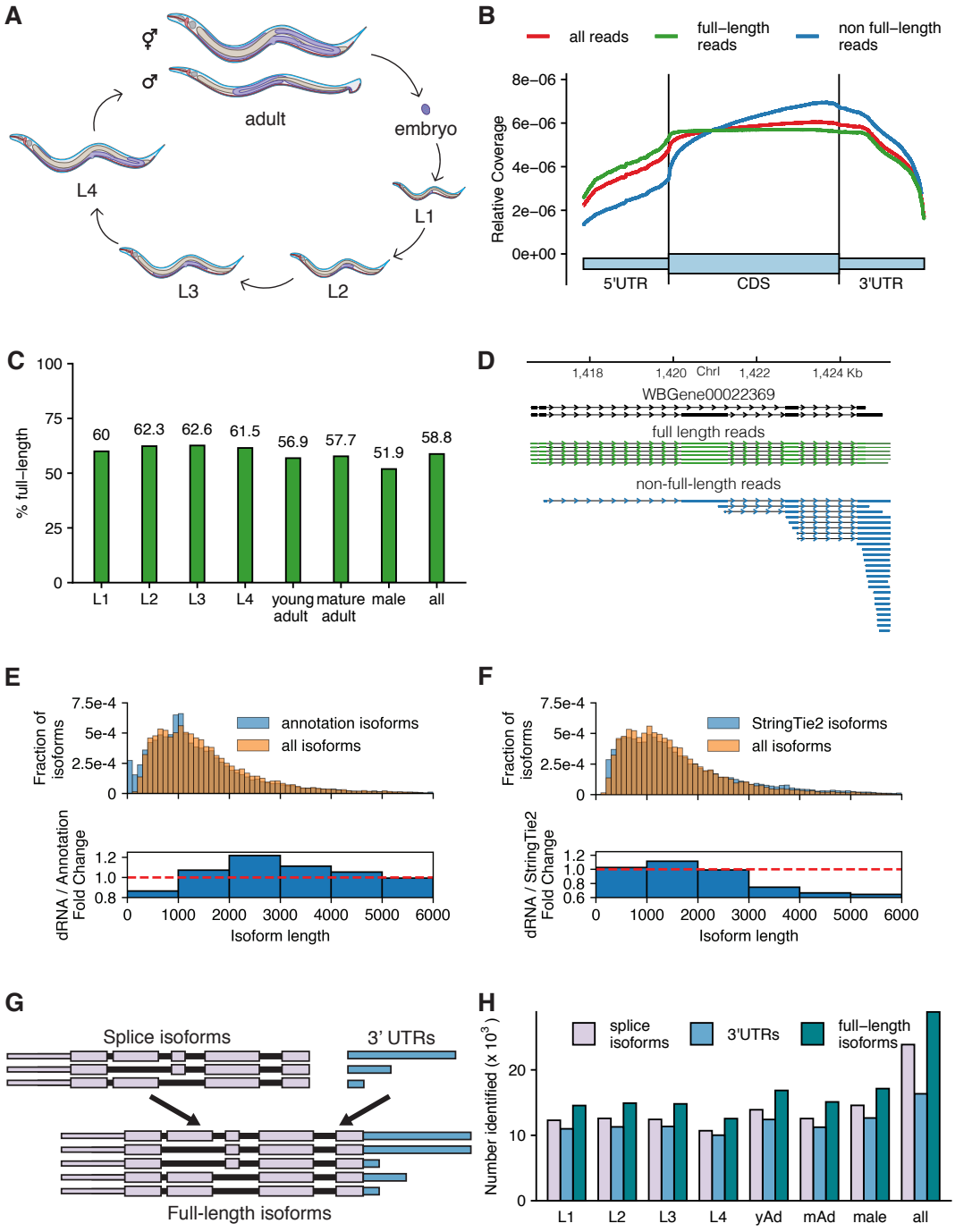
mean per base quality scores above 10 for each developmentally staged sample and median per base quality scores ranging between nine and 10 for each sample. Reads were aligned to the ce11 genome using minimap2, which successfully aligned 87.8% of our reads (Supplemental Table S2.2; (Li, 2018)). Median read lengths ranged between 573 and 687 for a given sample, while average read lengths were significantly longer, ranging from 739 to 934. Note that nanopore sequencing reads are currently unable to capture the last 10–15 bases proximal to the 5′ end because of the structure of the pore motor protein-RNA assembly as reported previously (Workman et al., 2019). The average percent reference identities of our alignments (as calculated by the NanoPack software suite [De Coster et al. (2018)]) ranged between 85.3% and 86.9% depending on the sample, suggesting an error rate of ~14%–15% in our data sets. This error rate and the loss of the 5′ -most 10–15 bases prevented us from examining splice leader-based trans splicing, a common RNA modification in *C. elegans* (see Section 2.11.1.1 in **Supplemental Material** 2.11). The average length of unique splice isoforms identified in our sequencing libraries was 1596 nt (Fig. 2.1E,F and discussed below), which was consistent with the annotated average length of transcripts in the WormBase WS265 annotation of the *C. elegans* transcriptome (1574 nt) (Lee et al., 2018; WormBase web site, 2018).

2.4.2 Identifying reads representing full-length transcripts

While the majority of our reads correspond to full-length transcripts (Fig. 2.1B,C), a substantial fraction of aligned reads failed to span the full length of an annotated coding sequence (31.8% of the unfiltered genome aligned reads);

these reads were predominantly truncated relative to annotated isoforms at their 5′ ends, resulting in a 3′ bias in coverage from our total reads (e.g., Fig. 2.1D). Including these reads in our downstream analysis would have artificially inflated the number of isoforms identified. Therefore, to make use of the long read lengths possible through dRNA-seq, reduce this 3′ bias, and eliminate the need to computationally reconstruct gene models, reads were filtered to ensure that only high-quality reads corresponding to full-length transcripts were considered (see **Methods** and Supplemental Fig. S2.1 for an outline of the entire analysis). Reads were considered full length in our filtering approach if they passed our complete filtering pipeline. Briefly, reads were discarded if they (1) contained large insertions or large 3′ softclips (i.e., bases at the end of a read that fail to align), (2) had no detectable poly(A) tail, (3) had 5′ ends that had insufficient evidence of corresponding to a transcription start site (TSS), (4) had a donor or acceptor splice site that could not be assigned to an annotated donor or acceptor splice site (i.e., a splice site

Figure 2.1 ***(following page)*****:** Overview of approach and sequencing of full-length isoforms. **(A)** Diagram of the *C. elegans* life cycle. **(B)** Plot of normalized coverage across the average coding gene with full-length (green) non-full-length (blue) and all reads (red) considered. **(C)** Percent of reads that passed filtering and were called full-length in each stage. **(D)** Example locus showing reads aligning to the WBGene00022369 locus (black). **(E)** Comparison of length distributions of isoforms present in the WormBase WS265 annotation, and splice isoforms identified by this study displayed as a density plot (top) and as the fold change of the densities (bottom). **(F)** As in E, comparison of length distribution of isoforms assembled by StringTie2 using Illumina based RNA-seq from across *C. elegans* development, and splice isoforms identified by this study. **(G)** Schematic defining "full-length isoform" as a combination of splice isoform and 3′ UTR isoform. **(H)** Number of splice, 3′UTR, and full-length isoforms observed across all stages. yAd = young adult, mAd = mature adult. Exact numbers can be found in Supplemental Table S2.3


A
adult
embryo
L1
L2
L3
L4
B
all reads
full–length reads
non full–length reads
Relative Coverage
8e–06
6e–06
4e–06
2e–06
0e+00
5'UTR
CDS
3'UTR
C
% full–length
100
75
50
25
0
60
62.3
62.6
61.5
56.9
57.7
51.9
58.8
L1
L2
L3
L4
young adult
mature adult
male
all
D
1,418
1,420
ChrI
1,422
1,424 Kb
WBGene00022369
full length reads
non-full-length reads
E
Fraction of isoforms
7.5e-4
5e-4
2.5e-4
0
annotation isoforms
all isoforms
dRNA / Annotation Fold Change
1.2
1.0
0.8
0
1000
2000
3000
4000
5000
6000
Isoform length
F
Fraction of isoforms
7.5e-4
5e-4
2.5e-4
0
StringTie2 isoforms
all isoforms
dRNA / StringTie2 Fold Change
1.2
1.0
0.8
0.6
0
1000
2000
3000
4000
5000
6000
Isoform length
G
Splice isoforms
3' UTRs
Full-length isoforms
H
Number identified (x 10^3)
splice isoforms
3'UTRs
full–length isoforms
20
10
0
L1
L2
L3
L4
yAd
mAd
male
all

not within 15 bp of an annotated splice site), (5) had retained introns, or (6) had 3′ ends that had insufficient evidence of corresponding to a bona fide polyadenylation site.

For 5′ end filtering criterion (step 3 above), we implemented a stringent 5′ filtering step that kept reads if their 5′ ends fell within -100 to +15 of an annotated transcription start site or were supported by 5′ SAGE data or high-throughput sequencing of RNA polymerase II initiation sites (Chen et al., 2013; Saito et al., 2013). For the 3′ end filtering (step 6 above), we kept all reads that overlapped with a stop codon in the WormBase WS265 GFF3 annotation (Lee et al., 2018; WormBase web site, 2018). For those reads that did not overlap with an annotated stop codon, we examined the read for canonical or alternative polyadenylation signals (PAS) up to 60 bp upstream of the putative 3′ end, as well as a predicted open reading frame (ORF) in the read with defined start and stop codons. We kept all reads that had both an ORF and a canonical or alternative PAS.

Our collection of filtering steps ensures that we keep only full-length transcripts with 5′ and 3′ ends that correspond to TSSs and polyadenylation sites, respectively, for further analysis. To determine the efficacy of this filtering approach, we made an aggregate plot of normalized coverage across the average coding gene (Fig. 2.1B). Supporting the validity of this approach, the reads that fail our filtering steps have an extreme 3′ bias, while those that pass the filtering steps do not have this 3′ bias in the total reads. Passing reads comprise the majority of reads in each data set (Fig. 2.1C). Combining all data sets, almost 2.9 million passing reads were obtained (Supplemental Table

S2.2 for a breakdown of reads remaining after each filtering step). Following read filtering, reads were assigned to the splice isoforms and 3'UTRs present in each developmental stage and across all stages, as described in **Methods, Section** 2.6.

2.4.3 Examining read and isoform length distributions

Part of the appeal of long-read RNA sequencing is the ability to capture full-length isoforms. However, as our library preparation is dependent on the poly(A) tail, 3'-biases may skew the resulting isoform length distribution and annotation of the transcriptome. To characterize this in our data sets, we plotted the length distribution of poly(A)-selected RNA from each of our stages as identified through TapeStation traces (Supplemental Fig. S2.2A). We then compared the TapeStation traces to the expected fluorescence signal based on the read lengths obtained from our nanopore sequencing experiments (determining expected fluorescence by weighting by the length of the reads) (Supplemental Fig. S2.2B). The expected fluorescence based on the sequencing read length distribution obtained is shorter than the distribution one would expect from an unbiased sequencing experiment based on TapeStation traces. However, we identified the two major peaks in our RNA length distributions as those corresponding to ribosomal subunits, indicating that oligo d(T) pull down of RNA failed to remove all ribosomal RNA from our samples. We next sought to determine if this read length bias was resulting in a shorter identified transcriptome on average compared to the existing transcriptome annotation (Fig. 2.1E; Supplemental Fig. S2.3A), and to transcriptome annotations assembled by StringTie2 (Kovaka et al., 2019) using (1) previously

collected Illumina RNA-seq data from across C. elegans development generated by the modENCODE Project (Hillier et al., 2009), (2) previous work from our laboratory (Weiser et al., 2017), and (3) the Albritton et al. (Albritton et al., 2014) study (Fig. 2.1F; Supplemental Fig. S2.3B). We find that despite the length difference observed between the TapeStation and nanopore read length, the length distribution of the unique isoforms we identify in our analyses are similar to the length distributions of the WormBase transcriptome annotation and transcriptome annotations produced by Illumina data and StringTie2. Taken together, these analyses indicate that our analysis pipeline mitigates the impact of any fragmentation-induced read length biases present in our sequencing and suggests that the full-length transcript isoforms we identify accurately reflect the structure and length of transcripts in the full-length *C. elegans* transcriptome.

2.4.4 Identifying the full-length transcriptome

The full-length single-molecule resolution of nanopore sequencing means that, unlike short-read sequencing, the full linear sequence of exons comprising a transcript and all of the associated splice junctions (i.e., the splice isoform) and the 3′-UTR isoform are captured together in a single read. This enables the identification of the "full-length transcriptome," the set of full-length isoforms (splice isoform + 3′-UTR isoform) observed together across all reads (Fig. 2.1G). When considered across all developmental stages and conditions, 28,858 full-length isoforms were identified, comprised of 23,865 unique splice isoforms and 16,342 unique 3'UTRs (Fig. 2.1H; Supplemental Table S2.3 for exact values). Over 12,000 full-length isoforms were identified in each stage.

Because 3'UTRs were only called if there were three or more reads supporting the putative cleavage site, not all splice isoforms have an associated 3'UTR called. Therefore, some full-length isoforms have no high-confidence 3'-UTR call and are, in effect, simply splice isoforms. This describes only a fraction (5583, 19%) of the full-length isoforms identified.

To determine if these data sets were at or approaching saturation in the number of full-length isoforms identified, reads were randomly subsampled and the number of full-length isoforms that had support from one or more reads in the subsampled set was determined. These values were then plotted, and the relationship between the number of reads considered and the number of full-length isoforms supported was examined. As expected, none of the developmentally staged data sets appears to be saturated (Supplemental Fig. S2.4A). We also examined the number of isoforms identified across all stages, which also does not yet appear to be saturated (Supplemental Fig. S2.4B).

The ability to resolve splice isoforms and 3'-UTR isoforms together at single-molecule resolution allows for identification of genes where the two features appear to be correlated. Few examples of significant correlations between splice isoform use and 3'-UTR isoform use were identified by Fisher's exact test after multiple hypothesis testing correction (Supplemental Table S2.4). This is possibly due to lack of coverage but more likely reflects an overall lack of coordination between splicing and polyadenylation site choice in *C. elegans*.

2.4.5 Quantifying genes and splice isoforms captured with full-length support

Less than half of the 30,133 isoforms with annotated introns in the WormBase WS265 annotation have full-length support (here, full-length support means that every annotated intron in the isoform is supported by the same cDNA or EST) (Fig. 2.2A; Lee et al. (2018); WormBase web site (2018)). By comparison, 17,245 splice isoforms (of the 30,133 with annotated introns in WormBase) across 13,400 genes had full-length support using our data, well above the 12,613 isoforms and 10,711 genes that have full-length support in the WormBase WS265 annotation (Fig. 2.2A). Comparing the genes and isoforms with full-length support in each data set, 4187 genes and 7247 isoforms were identified that did not previously have full-length support (Fig. 2.2B). This data set therefore significantly expands the number of *C. elegans* genes and isoforms supported by full-length data. To examine the changes of splice isoform usage in each developmental stage and across all stages, we plotted the number of previously annotated splice isoforms and genes observed in each stage (Fig. 2.2C; Supplemental Table S2.3). We found more than 10,000 previously annotated splice isoforms in each stage, with males having the most identified genes and splice isoforms of any individual stage despite having fewer reads after our filtering steps than most other stages (Supplemental Table S2.2). Combining across all stages, 20,413 previously annotated splice isoforms were observed. Most genes in our transcriptome data have only a single identified splice isoform, and the frequency of genes with a given number of isoforms decreases as the number of isoforms increases, consistent with the WS265

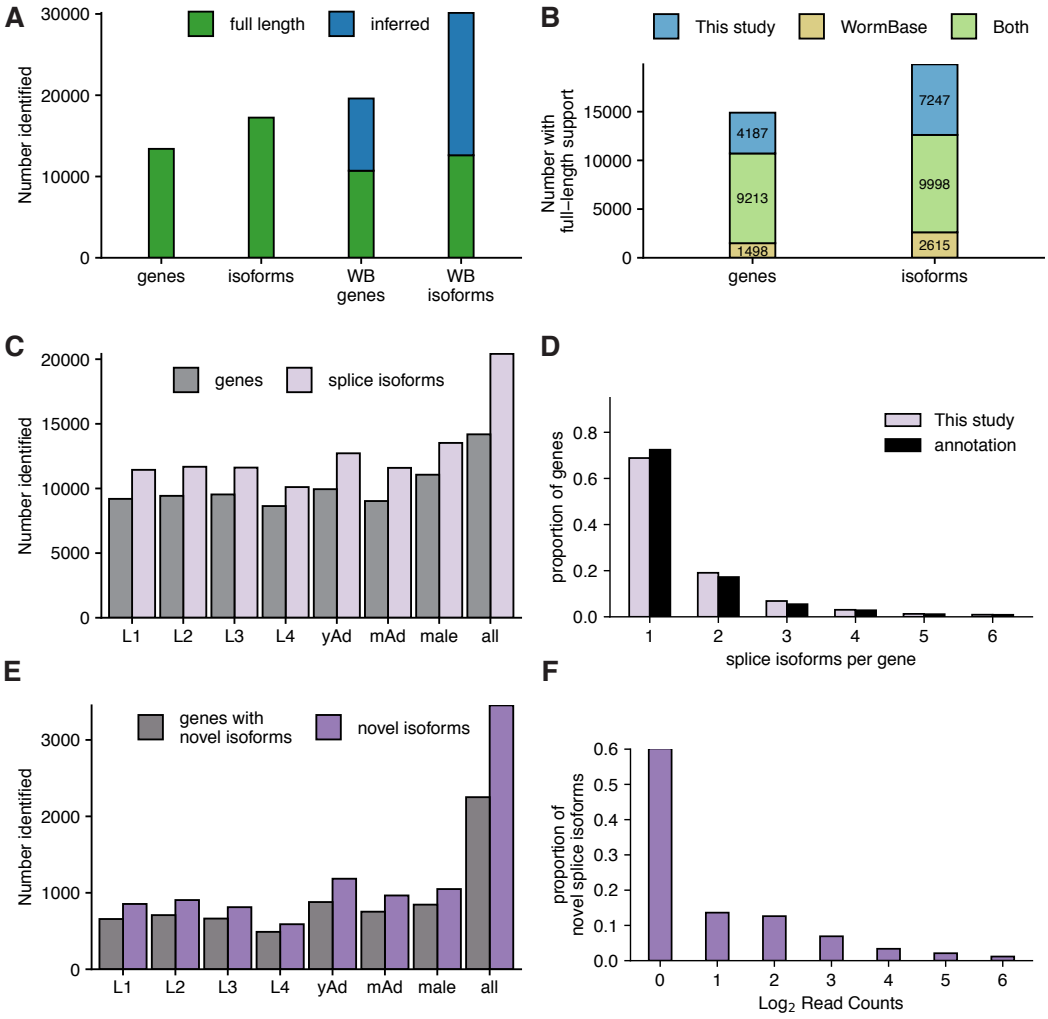


Figure 2.2: Capture of annotated and novel full-length splice isoforms. **(A)** Number of genes and isoforms captures with full-length support in our dataset (left) versus the WormBase (WB) annotation (right) (WormBase web site, release WS265, 2018; Lee et al. 2018) **(B)** Stacked bar graph showing overlap between isoforms and genes with full length support in our dataset and those with full-length support in the WormBase annotation. **(C)** Number of previously annotated splice isoforms and corresponding genes identified by our data across all stages yAd = young adult, mAd = mature adult. **(D)** Density plot showing the number of isoforms identified per gene across our full dataset and the WormBase annotation. **(E)** Number of novel isoforms and genes with novel isoforms identified across all stages. **(F)** Density plot showing the proportion of novel splice isoforms with a given number of reads supporting their structure.

annotation of the *C. elegans* transcriptome (Fig. 2.2D).

In addition to capturing previously annotated splice isoforms, part of the appeal of long-read single-molecule sequencing is the ability to detect novel splice isoforms. To test our ability to identify novel splice isoforms after stringent filtering and splice site correction steps, we searched for isoforms with a set of splice junctions not present in the WormBase WS265 annotation. Across all stages, 3452 novel splice isoforms were identified corresponding to 2251 genes (Fig. 2.2E; Supplemental Table S2.3). Of the novel splice isoforms, 1285 have novel splice junctions between previously annotated donor and acceptor splice sites, and 262 have novel exons. To determine the level of support for these novel isoforms, we generated a density plot showing the proportion of novel isoforms with a given number of reads supporting them (Fig. 2.2F). The majority of identified novel splice isoforms were identified with a single read supporting their structure; however, over 20% of novel isoforms had four or more reads supporting them, indicating that these are high-confidence novel isoforms.

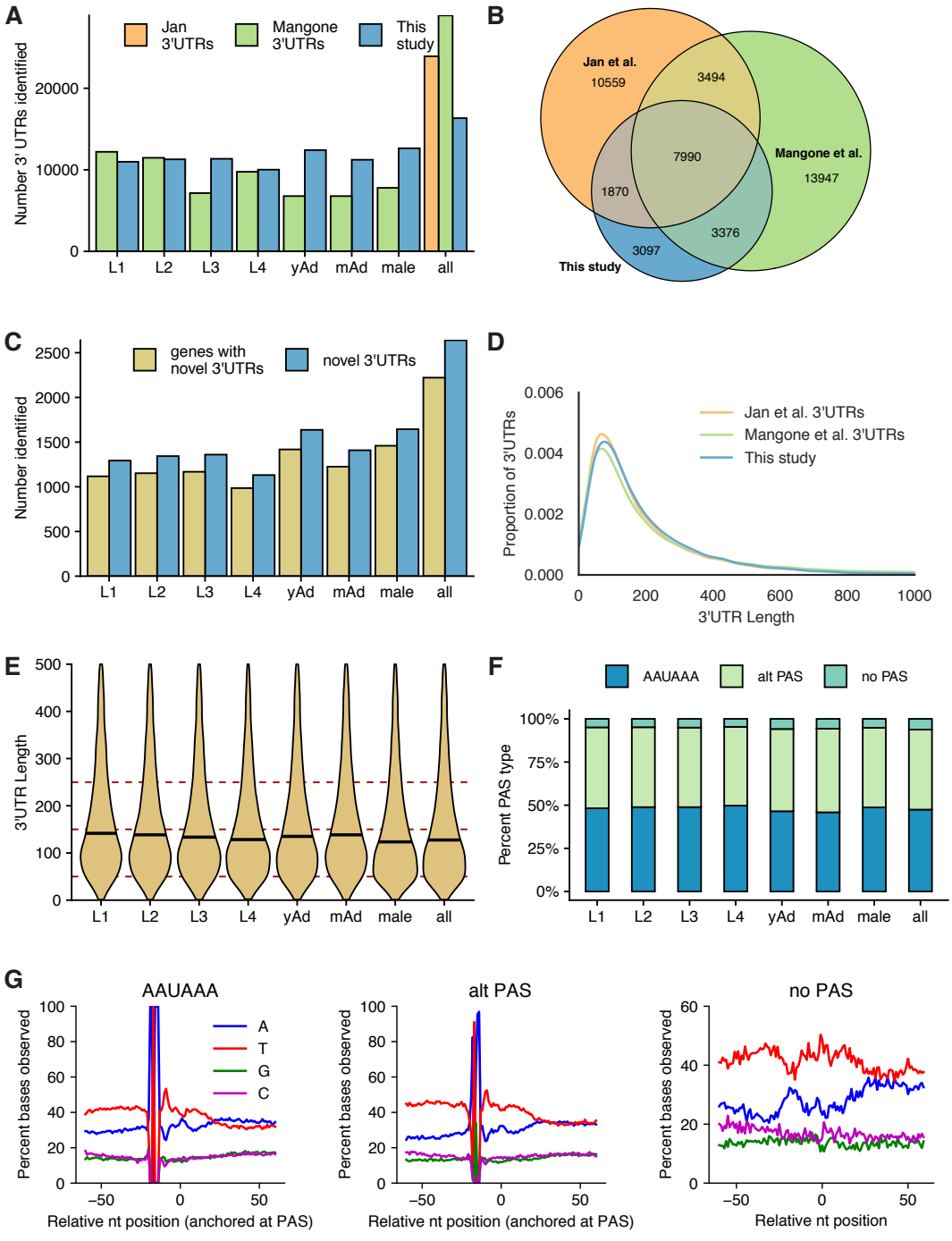
Finally, we sought to examine how many of our identified splice isoforms were predicted to be noncoding using the protein coding prediction algorithm CPAT trained on the *C. elegans* transcriptome annotation (Wang et al. 2013). Using this software, 1623 of our 23,865 splice isoforms appear to be noncoding (using a threshold of coding probability of 0.5 for defining the boundary between coding and noncoding isoforms, the IDs of which are listed in Supplemental Table S2.5).

2.4.6 Characterizing the identified 3'UTRome

Previous analyses of nanopore sequencing have largely centered on splice isoform identification and characterization while largely ignoring the 3'UTR. Because dRNA-seq relies on sequencing in the 3' to 5' direction of mRNAs isolated by their poly(A) tails, full-length sequences of 3'UTRs are preferentially captured. After adapter trimming, discarding reads with large 3' softclips, and realigning the 3' softclipped portions of the remaining reads, we identified putative poly(A) cleavage sites and predicted stop codons to define full-length 3'UTRs. Using this method, 16,342 unique 3'-UTR isoforms were identified, with over 10,000 3'UTRs identified in each stage (Supplemental Table S2.3). When splice structure in the 3'-UTR region is ignored to ease comparison with existing data sets (as described in Methods, Section 2.6), 16,333 3'UTRs are identified (Fig. 2.3A).

To determine the accuracy of our 3'-UTR calling, we compared the 3'UTRs identified by this method with those from previously published data sets

Figure 2.3 ***(following page)*****:** Properties of 3'UTRome **(A)** Number of 3'UTRs observed across all stages, as compared to Mangone et al. (Mangone et al. 2010) and Jan et al. (Jan et al. 2011). yAd = young adult, mAd = mature adult. **(B)** Venn diagram showing overlap between 3' UTRs identified in this study, Jan et al., and Mangone et al. **(C)** Number of novel 3'UTRs and genes with novel 3'UTRs identified in each stage and across all stages **(D)** Kernel density estimate plot of 3'UTR lengths from this study, Jan et al. and Mangone et al. **(E)** Violin plots showing 3'UTR length distributions across all stages. Horizontal black lines show the median of each stage. **(F)** Stacked bar chart showing percentage of UTRs with the specified polyadenylation signal (PAS) across all stages. **(G)** Nucleotide distributions around putative PAS sites and putative cleavage sites. Canonical PAS (AAUAAA) and alternative PAS (alt PAS) distributions are anchored with the putative PAS hexamer at -19 nucleotides. The distribution of UTRs with no PAS is anchored with the putative cleavage site at 0


A
Jan 3'UTRs
Mangone 3'UTRs
This study
Number 3' UTRs identified
0
10000
20000
L1
L2
L3
L4
yAd
mAd
male
all
B
Jan et al.
10559
3494
Mangone et al.
13947
7990
1870
3376
3097
This study
C
genes with novel 3'UTRs
novel 3'UTRs
Number identified
0
500
1000
1500
2000
2500
D
Jan et al. 3'UTRs
Mangone et al. 3'UTRs
This study
Proportion of 3'UTRs
0.000
0.002
0.004
0.006
0
200
400
600
800
1000
3'UTR Length
E
3'UTR Length
0
100
200
300
400
500
F
AAUAAA
alt PAS
no PAS
Percent PAS type
0%
25%
50%
75%
100%
G
AAUAAA
alt PAS
no PAS
A
T
G
C
Percent bases observed
Relative nt position (anchored at PAS)
Relative nt position

(including 3P-Seq and 3′ RACE data) generated in *C. elegans* (Mangone et al., 2010; Jan et al., 2011). Of our identified 3′UTRs, 81.0% overlap with one or more of these 3′-UTR data sets (Fig. 2.3B). In addition, we identified 2640 novel 3′UTRs that do not fall within 10 bp of existing 3′UTRs or WormBase 3′-UTR annotations (Fig. 2.3C). The 3′-UTR length distribution in our data was nearly identical to those observed by Jan et al. and Mangone et al. (Fig. 2.3D). In agreement with Mangone et al., our 3′-UTR length distributions change over developmental stages, progressively decreasing from L1 through L4, and are shorter in males than in hermaphroditic adults (Fig. 2.3E). The 3′-UTR length distributions in adult stages were slightly longer than the length distribution of L4 3′UTRs in our data sets, in contrast to Mangone et al. (Mangone et al., 2010), which showed that adult 3′UTRs had a slightly shorter average 3′-UTR length than L4.

Given that the lengths of 3′UTRs change during development, we investigated whether PAS usage might also vary across time. We compared the frequency of canonical PAS usage (defined by the motif AAUAAA), alternative PAS usage (defined by a subset of hexamers with a 1- or 2-nt difference from AAUAAA [see Methods, Section 2.6.14]), and sites with no defined PAS. Frequency of canonical and alternative PAS usage was quite consistent between adjacent developmental stages, although by χ^2 tests, there were statistically significant differences in overall PAS usage between the L4 and young adult stage as well as between hermaphroditic young adults and males (Fig. 2.3F). Given that distribution of canonical and alternative PAS usage is consistent across the larval stages, where a significant shift in 3′-UTR length distributions

occurs, this suggests that 3′-UTR length changes over development are largely independent of PAS usage.

As a final metric for the accuracy of this 3′UTRome, we plotted nucleotide distributions in windows around identified PAS sites and around putative cleavage sites (Fig. 2.3G). This largely agrees with previously published nucleotide distributions in windows around identified PAS sites (Mangone et al., 2010). These distributions are AT-rich, with a peak in T frequencies just 3′ from the PAS site. It is possible that 3′UTRs identified by our method were inaccurate and broadly distributed around true cleavage sites, and by anchoring nucleotide distributions with putative PAS sites at -19 nt, the impact of these errors was eliminated. To test this possibility, we generated a density plot of the offsets of identified PAS sites from putative cleavage sites identified by our method and found that these offsets were enriched close to the canonical -19 nt from putative cleavage sites, indicating cleavage site calls from this method are accurate within a few base pairs (Supplemental Fig. S2.5A,B). At 3′-UTR sites without a putative PAS identified, the nucleotide distribution observed lacks the enrichment of As in a window around the cleavage site noted in Mangone et al. (2010). Our method may be capturing a different set of 3′UTRs with no PAS than the Mangone et al. data set. Supporting this possibility, only 28% of the no-PAS 3′UTRs in our data set overlap with a Mangone et al. 3′UTR, as compared with 71% of canonical PAS and 64% of alternative PAS 3′UTRs in our data (Supplemental Fig. S2.5C). In addition, no-PAS 3′UTRs that do overlap with a Mangone et al. 3′UTR have a different nucleotide distribution than the no-PAS Mangone et al. 3′UTRs in general

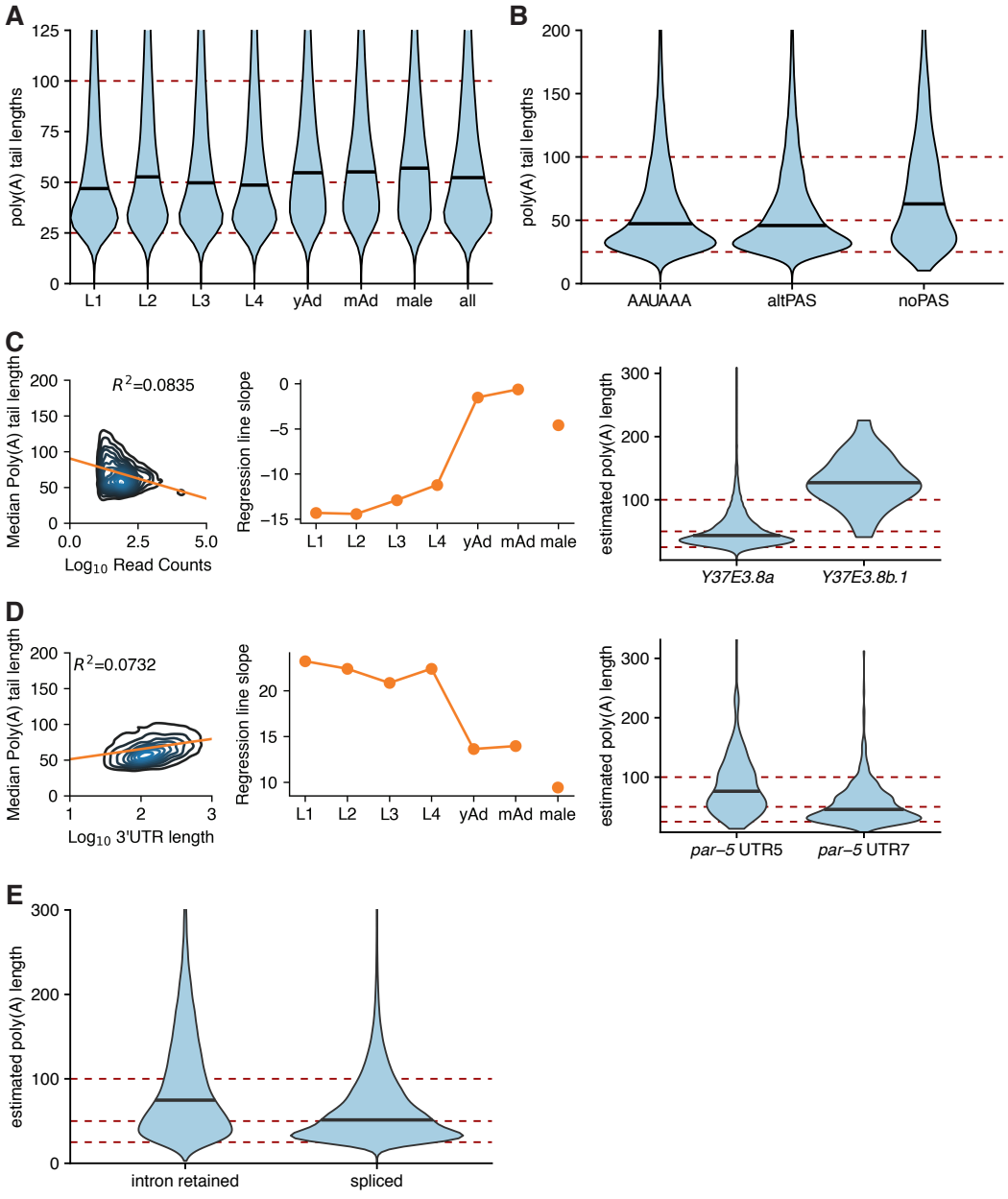
(Supplemental Fig. S2.5D; Mangone et al. (2010)).

2.4.7 Properties of poly(A) tail lengths

Poly(A) tails are known regulators of translation and transcript stability. However, profiling of poly(A) tail lengths at the transcriptome-wide level using short-read sequencing is a relatively recent advance in the field (Chang et al., 2014; Subtelny et al., 2014; Lim et al., 2016). We have previously shown that, using a trained hidden Markov model, one can estimate the poly(A) tail length of dRNA-seq reads using nanopolish (Workman et al., 2019). We performed these estimations on our current data, providing a developmentally resolved poly(A) profiling data set.

Global poly(A) tail length distributions are dynamic in the developing *Drosophila melanogaster* oocyte and embryo (Lim et al., 2016). To determine if there

Figure 2.4 *(following page)*: Properties of poly(A) tail length **(A)** Violin plot of poly(A) tail length distributions across development. Horizontal black lines show the median of each stage. yAd = young adult, mAd = mature adult. **(B)** Poly(A) tail length distributions separated by the PAS type of the associated reads for reads corresponding to isoforms predicted to be coding. **(C)** (*left*) Density plot showing correlation between poly(A) tail length and expression level by plotting median poly(A) tail length for each isoform versus the log of the expression level of that isoform (across all stages). Linear regression plotted in orange. (*middle*) Slope of linear regressions performed on median poly(A) tail length versus expression level data across developmental stage. (*right*) Example locus illustrating relationship between poly(A) tail length and expression level *Y37E3.8b.1* is lower expressed than *Y37E3.8a* with a longer poly(A) tail length distribution **(D)** (*left, middle*) As in the left and middle panels of **(C)**, but instead plotting median poly(A) tail length versus the log of the 3'UTR length. (*right*) Example locus illustrating relationship between 3'UTR length and poly(A) tail length; par-5 UTR 5 is longer than par-5 UTR 7, and has a longer poly(A) tail length distribution. **(E)** Violin plots showing poly(A) tail length distributions in fully spliced versus intron retention transcripts.


A
poly(A) tail lengths
125
100
75
50
25
0
L1
L2
L3
L4
yAd
mAd
male
all
B
poly(A) tail lengths
200
150
100
50
0
AAUAAA
altPAS
noPAS
C
Median Poly(A) tail length
R^2=0.0835
Log10 Read Counts
Regression line slope
L1 L2 L3 L4 yAd mAd male
estimated poly(A) length
Y37E3.8a
Y37E3.8b.1
D
Median Poly(A) tail length
R^2=0.0732
Log10 3'UTR length
Regression line slope
L1 L2 L3 L4 yAd mAd male
estimated poly(A) length
par–5 UTR5
par–5 UTR7
E
estimated poly(A) length
intron retained
spliced

were comparable shifts in our poly(A) tail length distributions, we examined poly(A) tail lengths across the developmental stages in *C. elegans*. The poly(A) tail length distributions display only modest fluctuations, ranging from median values of 49 nt (L1) to 54 nt (L2) during larval development, although these shifts were considered to be statistically significant by Kolmogorov–Smirnov and Mann–Whitney U tests (Fig. 2.4A). However, length distribution in all adult stages (young and mature hermaphrodites and males) are consistently longer than in the larval stages, with a median length of 58 nt in adults compared to an aggregate median length of 52 nt across all larval stages ($P < 2.2 \times 10\text{-}16$ by Kolmogorov–Smirnov and Mann–Whitney U tests). These data suggest that the most significant regulation of poly(A) tail lengths occurs between larval and adult stages during development.

As a means of confirming the validity of our poly(A) tail length profiling approach, we compared our poly(A) estimates from the L4 stage with previously published poly(A) measurements from the L4 stage of *C. elegans* from mTAILseq (Lima et al., 2017). The length scale distributions of our L4 data and the Lima et al. data set are quite similar, as both have peaks around 30–40 nt and extended toward the longer tail length range (Supplemental Fig. S2.6A). However, we did not identify the shoulder peaks present in the Lima et al. data set (Lima et al., 2017).

An advantage of profiling poly(A) tail lengths with dRNA-seq versus short-read sequencing is that poly(A) tail lengths are directly coupled to information about the splice isoforms and 3′-UTR isoforms of the associated read. This allows comparisons and correlations between poly(A) tail lengths and aspects

of transcript structure. One possible driver of differences in poly(A) tail lengths between reads could be that poly(A) tail length distributions may vary depending on whether the associated 3'UTR has a canonical PAS site. To test this possibility, we plotted poly(A) tail length distributions versus PAS type (i.e., canonical AAUAAA, alternative PAS, and no PAS) for reads from the L1 stage corresponding to isoforms predicted to be coding (based on the CPAT prediction algorithm [Fig. 2.4B; Wang et al. (2013)]). We find that all PAS types are significantly different from one another by Kolmogorov–Smirnov and Mann–Whitney U tests ($P < 2.2 \times 10\text{-}16$), and 3'UTRs with no PAS have longer poly(A) tail lengths, on average, than poly(A) tails associated with either canonical and alternative PAS, with a median poly(A) tail length of 62 nt for 3'UTRs with no PAS, 46 nt for 3'UTRs with alternative PAS, and 48 nt for 3'UTRs with canonical PAS.

It has been reported that median poly(A) tail length and expression level are anticorrelated, such that highly expressed genes generally have shorter median poly(A) tail lengths (Lima et al. 2017; Legnini et al. 2019). To determine if this relationship holds in our data sets, we plotted the log of the number of reads supporting a given isoform versus the median poly(A) tail length for that isoform for transcripts with 10 or more reads supporting them (Fig. 2.4C, left panel; Supplemental Fig. S2.6B). A similar inverse correlation between median poly(A) tail length and number of reads supporting that isoform was observed in the L1 to L4 stages and when all stages were pooled (Supplemental Fig. S2.6B). For example, the a isoform of the Y37E3.8 gene (Y37E3.8a) is expressed much more than the b.1 isoform (Y37E3.8b.1; 13,299 reads vs. 26 reads) and has

a significantly shorter poly(A) tail length distribution than the b.1 isoform (Fig. 2.4C, right panel). However, this correlation explains only a small fraction of the overall variation in the data, with the maximum R2 value of 0.1242. In the adult stages (both males and hermaphrodites), the slope of the regression lines between median poly(A) tail length and expression level were much shallower, and the corresponding R2 values were much weaker, with R2 values ranging from 0.0003 to 0.0122 (Fig. 2.4C, middle panel; Supplemental Fig. S2.6B). These results suggest that the inverse relationship between poly(A) length and expression level may vary depending on the developmental stage.

A recent study using FLAM-seq, a Pacific Biosciences (PacBio) sequencing method that also captures poly(A) tails and fulllength transcripts, demonstrated that poly(A) tail length and 3′-UTR length were positively correlated (Legnini et al., 2019). Examining poly(A) tail length and 3′-UTR lengths across all reads in our data, we also identify this same relationship (Fig. 2.4D, left panel). For example, the longer par-5 3′-UTR isoform (termed 3′UTR 5; 486 nt) also has a longer poly(A) tail (median length 74 nt) versus the shorter par-5 3′-UTR isoform (3′UTR 7; 51 nt) with a shorter poly(A) tail length distribution (median length 45 nt) (Fig. 2.4D, right panel). However, the overall strength of this relationship also varies between developmental stages, and the slopes of the regression lines (and the corresponding R2 values) are smaller in adult stages than in larval stages (Fig. 2.4D, middle panel; Supplemental Fig. S2.6C).Finally, we examined the poly(A) tail length distributions between transcripts that are fully spliced versus those with retained introns (Fig. 2.4E). We previously showed in the human cell line GM12878 that intron retention

correlates with transcripts with longer poly(A) tails (Workman et al. 2019). In our *C. elegans* data sets, we also found a positive correlation between intron retention and poly(A) tail length distributions by Kolmogorov–Smirnov and Mann–Whitney U tests, suggesting a conserved mechanism whereby nuclear transcripts possess longer poly(A) tails and supporting a model proposed by Lima et al. (2017) in which poly(A) tails may be subject to post-transcriptional processing by deadenylation once exported into the cytoplasm.

2.4.8 A public resource for full-length isoform information

To make our transcriptome data set accessible to the research community, we have created a public custom track hub (https://bx.bio.jhu.edu/track-hubs/dRNAseq/hub.txt). This track hub contains the full-length filtered and nonfiltered reads from each developmental stage, as well as the full-length isoforms supported across all stages at each locus. To ease access to this track hub, we registered it with the Track Hub Registry (https://trackhubregistry.org). Users can therefore easily load this track hub in Ensembl based genome browser (Zerbino et al., 2018) by searching public track hubs for "ce11 staged dRNAseq". As a proof of the utility of this track hub, we loaded the track hub in the Ensembl Genome Browser and searched for lin-14, a gene with a well-studied 3'UTR that is subject to regulation by the lin-4 microRNA (Wightman et al., 1991; Lee et al., 1993) but whose 3'UTR is not currently annotated in the WormBase WS265 annotation (Lee et al., 2018; WormBase web site, 2018). In our data set, we identified the lin-14 3'UTR, as well as its splice isoforms, including multiple novel splice isoforms (Fig. 2.5A, "observed isoforms" track). As another example of the utility of this track hub, we searched for the locus

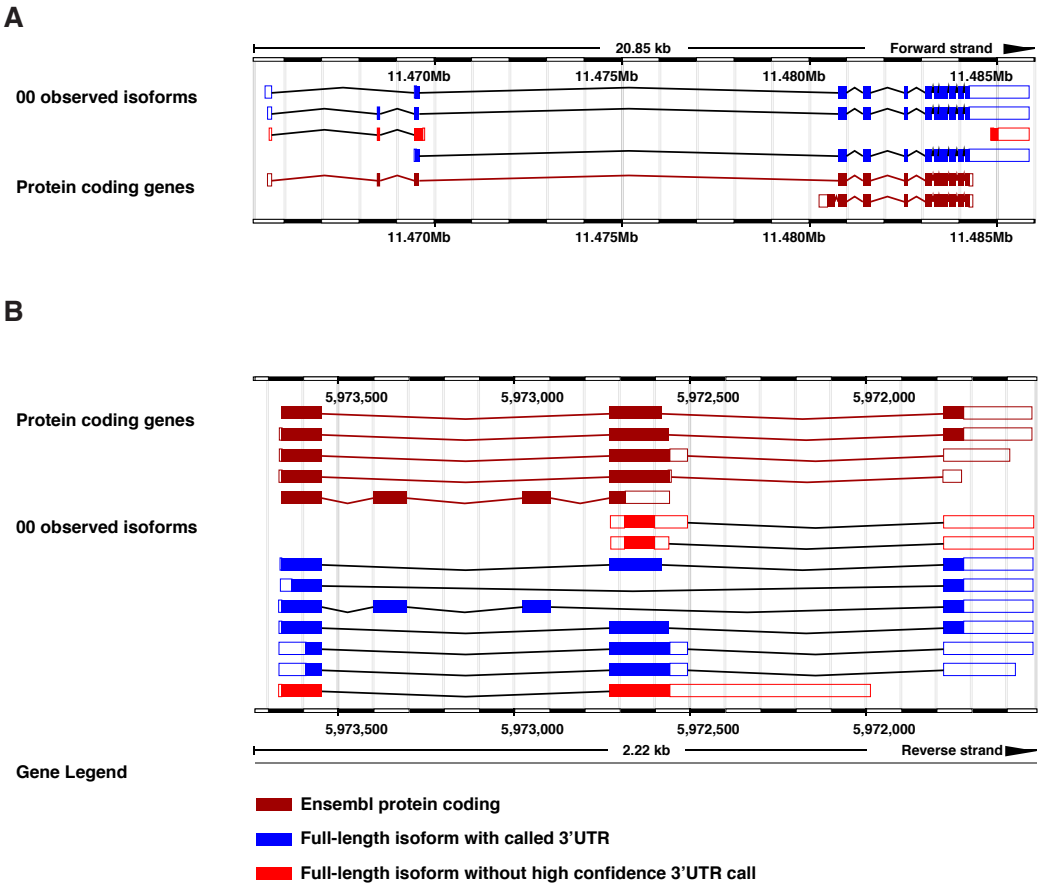


Figure 2.5: The *lin-14* **(A)**, or *mlp-1* **(B)** locus in the Ensembl genome browser including our custom track hub. Blue isoforms are full-length isoforms with an associated 3'UTR called, red isoforms have no high confidence 3'UTR called. Burgundy isoforms are protein coding models imported from WormBase.

mlp-1, a gene with multiple splice and 3'-UTR isoforms identified, including multiple novel splice isoforms (isoforms 1, 2, 4, 5, and 9 of the observed isoform track in Fig. 2.5B). These examples highlight possible uses of this resource by the research community to query currently unannotated 3'UTRs and splice isoforms.

2.5 Discussion

Despite years of study, our understanding of the *C. elegans* transcriptome remains incomplete. Although studies have been performed profiling transcription start sites, splicing in both *cis* and *trans*, 3'-UTR isoforms, poly(A) tail lengths, RNA base modifications, and gene and isoform expression levels, the short read lengths intrinsic to the prevailing technologies have limited the examination to one or two of these features at a time (Hillier et al., 2009; Mangone et al., 2010; Jan et al., 2011; Saito et al., 2013; Zhao et al., 2015; Lima et al., 2017; Tourasse et al., 2017; West et al., 2018; Packer et al., 2019). Even within these data sets, short read lengths and reliance on PCR amplification eliminate single-molecule resolution and make correlation of distant features within transcripts impossible. Although our study focuses primarily on splice isoforms, 3'-UTR isoforms, and poly(A) tail lengths due to current limitations of nanopore sequencing technologies, in principle, modified approaches to dRNA-seq would be capable of capturing all of the above features at a single-molecule level.

Nanopore sequencing therefore poses both a unique set of opportunities and challenges that must be addressed in any analysis pipeline. The dRNA-seq pipeline FLAIR (full-length alternative isoform analysis of RNA) utilizes a hybrid sequencing approach in which matched short-read sequencing is used to correct splice junctions in reads, and reads are clustered together into splice isoforms if they share a common set of splice junctions (Tang et al., 2018).

We utilized an approach similar to that used by FLAIR, in which reads are corrected, in our case by an existing annotation, and clustered together by splice isoform. Our approach differs from FLAIR in several ways, including a series of filtering steps that reduces the impact of 3′ bias in our reads and allows us to consider only full-length transcripts. A recent publication examining the utility of dRNA-seq and cDNA nanopore sequencing to generate transcriptome annotations independently revealed that many nanopore sequencing reads fail to span the full-length of annotated transcript isoforms, highlighting the need for analysis pipelines that take the possibility of 5′ truncations into account in isoform identification (Soneson et al., 2019). Our full-length filtering approach partially addresses this concern, although, as noted by Soneson and colleagues, doing so reduces the number of usable reads and likely impacts the quantitative nature of our data. A possible experimental approach to solving this problem could involve ligating a set of known nucleotides to the 5′ end of RNA transcripts after a decapping reaction, allowing for selection of full-length transcripts by filtering for reads flanked by signals corresponding to a poly(A) tail and the 5′ ligated product. This approach would incidentally also address the known problem of the 10–15 nt at the 5′ end of each strand that are unable to be read by nanopore sequencing methods (Workman et al., 2019).

Also distinguishing our approach from FLAIR is a novel means of calling 3′UTRs used in the generation of transcriptome annotations. We identify 3′-UTR structures with a standard dRNA-seq library preparation protocol, meaning that, in principle, any dRNA-seq experiment can be used to identify

3'UTRs using our method. The implications of this are potentially wide-reaching, as experiments once used for comparative analysis of splice isoforms between conditions may now also be used in comparative analysis of 3'-UTR isoforms.

By combining our 3'-UTR and splice isoform calls, we identified over 28,000 full-length transcript isoforms. It is likely that increased depth and additional sequencing of other developmental stages such as embryos and the stress-induced dauer stage would further increase the number of genes and isoforms identified, bringing this data set closer to capturing the theoretical complete C. elegans transcriptome.

The ability to estimate poly(A) tail lengths for each read is another advantage of dRNA-seq. Supporting the validity of our poly(A) profiling approach, the length distribution of the poly(A) tail length estimates we obtain in the L4 stage are quite similar to the distribution in the L4 stage reported by Lima et al., a study utilizing mTAIL-seq (Lima et al., 2017). Coupling of poly(A) tail lengths to aspects of 3'-UTR structure and splice isoform allowed us to identify relationships between putative PAS sites and intron retention transcripts to poly(A) tail lengths. The relationship between PAS sites and poly(A) tail lengths is a result that indicates there may be differential deposition or regulation of poly(A) tail length based on the presence or absence of an upstream PAS sequence. Longer poly(A) tails in intron retention transcripts could be indicative of partially processed RNAs retained in the nucleus, as nuclear RNAs would be shielded from cytoplasmic deadenylation. Neither of these relationships could have been discovered by short-read sequencing

of poly(A) tails, demonstrating the efficacy of full-length single-molecule sequencing.

One discovery of developmentally resolved poly(A) tail length profiling was the difference in features of poly(A) tail lengths between larval and adult stages. Overall, poly(A) tail length distributions were longer in adult stages than in larval stages, and the strength of previously reported correlations between poly(A) tail lengths and expression level and poly(A) tail lengths and 3'- UTR lengths were weaker in adult stages than larval stages. One possible explanation for these differences is the development of a functional germline in adult stages. In hermaphrodites, the cytoplasmic polyadenylases gld-2 and gld-4 are known to be active in the germline (Suh et al., 2006; Schmid et al., 2009; Millonigg et al., 2014; Nousch et al., 2017). Given the relative size of the *C. elegans* germline, it is possible that activity of such cytoplasmic poly(A) polymerases may influence global poly(A) tail length distributions.

Finally, we have created a custom track hub for exploration of this data set by independent researchers. By making this data easily accessible, we hope to provide *C. elegans* researchers with information related to their genes of interest, providing a resource to identify what isoforms have full-length support in any given developmental stage, and across all stages, as well as the structure of any 3'UTRs that we identify. Given that our data set provides support for over 7000 isoforms previously lacking full-length support and over 23,000 splice isoforms overall and given that most isoforms have an associated 3'UTR called, this will be a resource for the *C. elegans* research community. Overall,

we have demonstrated the utility of nanopore sequencing in providing support for full-length transcripts, annotating putative 3'UTRs, and interrogating poly(A) tail lengths.

2.6 Methods

2.6.1 *C. elegans* strains, maintenance, and collection

C. elegans N2 worms were grown and maintained under standard laboratory conditions on NGM plates seeded with *Escherichia coli* OP50 (Stiernagle, 2006). Samples for RNA analysis were synchronized by hypochlorite treatment and overnight hatching in M9 buffer. They were plated as starved L1 diapause worms at 25°C and staged by pharyngeal pumping. L2, L3, L4, and young adult (YA) worms were collected ~2 h postlethargus. L1 worms were collected 4 h after plating. Mature adults were collected ~10 h postL4/YA transition. CB1489 [him-8(e1489)IV] adult males were enriched by filtering through 35-µm mesh.

2.6.2 RNA extraction

Total RNA isolation was performed using TRI Reagent (Ambion) following the vendor's protocol, with the following alterations: Three rounds of freeze/thaw lysis were conducted prior to the addition of BCP; RNA was precipitated in isopropanol supplemented with glycogen for 1 h at -80°C; RNA was pelleted by centrifugation at 4°C for 30 min at 20,000g; the pellet was washed three times in 70% ethanol; the pellet was resuspended in water.

2.6.3 Library preparation and sequencing

Approximately 20-µg aliquots of total RNA were diluted to a total volume of 100 µL in nuclease-free water and poly(A)-selected using NEXTflex Poly(A) Beads (BIOO Scientific, Cat#NOVA-512980). Up to 600 ng of the resulting poly(A) RNA was separately aliquoted for library generation. Any excess poly(A)-selected RNA was stored at -80°C. Biological poly(A) RNA and a synthetic control (Lexogen SIRV Set 3, 2.5 ng) were prepared for nanopore direct RNA sequencing generally following the Oxford Nanopore Technologies (ONT) SQK-RNA001 kit protocol, including the optional reverse tran scription step recommended by ONT. One difference from the standard ONT protocol was use of SuperScript IV (Thermo Fisher Scientific) for reverse transcription. RNA sequencing on the GridION platform was performed using ONT R9.4 flow cells and the standard MinKNOW protocol script (NC_48Hr_sequencing_FLOMIN106_SQKRNA001).

2.6.4 Preprocessing and alignments

Reads were base-called and trimmed of adapter sequences using Poreplex version 0.3.1 (running Albacore version 2.3.1) with the following parameters: -p 24 –trim-adapter –basecall (https://github.com/hyeshik/poreplex). For each of our samples, reads were aligned to the WBcel235 ce11 genome using minimap2 version 2.14-r883 (Li, 2018). Genomic alignments were run with the following parameters: -ax splice -k14 -uf –secondary = no -G 25000 -t 24. The resulting SAM files were converted to BAM format using SAMtools view with parameters: -b -F 2048 (Li et al., 2009).

2.6.5 Read filtering

Our first filtering step involved removing reads aligning to the genome with large insertions (>20 bp) and large 3′ softclips (>20 bp) that could be the result of not properly aligning internal or 3′ exons, respectively. This filtering step ensures that novel isoforms identified in downstream scripts are not false positives resulting from poor alignments.

Following this, reads were filtered based on their QC tags from the poly(A) estimation module of the program nanopolish (Workman et al., 2019). Reads were removed from consideration if they had QC tags "READ_FAILED_LOAD", "SUFFCLIP", or "NOREGION". This was meant to remove reads without a detectable poly(A) tail signal, to prevent inclusion of reads with truncated 3′ ends.

Next, for the purposes of better identifying 3′-UTR isoforms in downstream analysis, 3′ soft-clips were realigned using a semiglobal aligner with affine gap penalties anchored at the 3′ end of the original alignment. This resulted in more uniform 3′ ends of alignment. The resulting realigned reads were converted to BED12 format using the BEDTools bamtobed function (version 2.27.1) (Quinlan and Hall, 2010; Quinlan, 2014).

Reads were then filtered to ensure that their 5′ ends reflected a bona fide TSS. This filter selected for reads with 5′ ends either within -100 to +15 bp of an annotated 5′ end of a transcript in the WormBase WS265 GFF3 file, within 5′ SAGE, or RNA polymerase II initiation clusters from Saito et al. and Chen et al., or within 10 bp of a called TSS from the same data (Chen et al., 2013;

Saito et al., 2013). Note that many of the 5′ ends of transcripts from the WS265 annotation do not reflect the true TSS of the gene but the end of outrons that are spliced out of the mature transcript at sites of trans splicing. Our approach, therefore, makes extensive fuse of trans splicing acceptor sites when defining full-length transcripts.

To account for errors in splice junction alignments, we used the WormBase WS265 GFF3 annotation to define canonical donor and acceptor splice sites and assigned each donor and acceptor splice site in our reads to a canonical splice site. Noncanonical donor and acceptor splice sites in our reads that fell within 15 bp of a canonical site were assigned to that site. Reads that contained noncanonical donor and acceptor splice sites that were not within 15 bp of a canonical site were discarded and not considered for the purposes of defining splice isoforms or UTRs. In addition, reads were thrown out if splice junctions in that read corresponded to annotated splice junctions from more than one gene. This allowed us to assign each spliced read to a gene based on its correspondence to annotated donor and acceptor splice sites. Reads were assigned to splice isoforms in a similar manner (however, some of these assignments were ambiguous when two annotated isoforms were comprised of the same sets of splice junctions). For non-spliced reads, we assigned gene IDs based on overlap with single exongenes present in the annotation.

Next, we separated reads that had exons that span the full length of any intron in the annotation that is not fully spanned by an exon in the annotation. We do this to separately consider intron retention transcripts when defining putative isoforms, as we believe these reads to be nuclear RNA that has not

been fully processed, which, if included, would artificially inflate the number of identified isoforms. Intron retention reads are only considered in analysis of poly(A) tail length distributions, in the comparison of poly(A) tail length distributions in fully spliced versus intron retention transcripts, and as a separately reported track in our custom Track Hub.

Finally, we implemented a filter to ensure we could be highly confident in our 3′ ends. This filter first keeps all reads that overlap with an annotated stop codon (extracted from the WormBase WS265 GFF3 annotation file into a stop_codons.bed file) (as determined using BEDTools intersect with the flags -u -s -split) (Lee et al., 2018; WormBase web site, 2018). Of the reads that do not overlap with an annotated stop codon, we search the 3′ end of the read for a putative canonical or alternative PAS, and we perform open reading frame predictions to determine if the read has a predicted open reading frame with a defined start and stop codon. If the read has both a putative PAS and a putative ORF, the read is kept; otherwise, the read is not considered in downstream analyses.

Reads were excluded from consideration in 3′-UTR calling (but not splice isoform calling) if their original minimap2 alignments had 3′ softclips larger than 10-nt long. This exclusion prevented reads with 3′ ends that failed to align well from being considered and reduced the variation in considered 3′ alignment ends significantly.

2.6.6 Splice isoform identification

After the intron retention filter and before our 3′ filter, we extracted the sequences from the ce11 WBcel235 genome corresponding to each aligned read using the getfasta function of the program BEDTools with the following flags: -s -split -bedOut (Quinlan, 2014). After completing the 3′ filtering step, we then clustered reads (and their associated sequences) together into putative isoforms if the reads shared a common set of splice junctions. This resulted in reads clustered by splice isoform. For each of these sets of reads corresponding to splice isoforms, we selected the longest read. From this read, we extracted information about the isoform including putative coding sequence by identifying the longest open reading frame (with both start and stop codons) present in the read's associated sequence. This allowed us to define putative start and stop codons.

Splice isoforms were called as novel if they contained a set of splice junctions not previously annotated in the reference. To deal with the possibility of 5′ truncated reads artificially inflating our novel isoform counts, we considered all possible 5′ truncations of previously annotated transcripts in the WormBase WS265 annotation file when defining our reference.

2.6.7 Generating an Illumina-based transcriptome annotation with StringTie2

We utilized Illumina RNA-seq reads from across C. elegans development (namely L1–L4, young adult, mature adults, and males) (for accession numbers, see Supplemental Table S2.6). These libraries were generated by the

modENCODE Project (Hillier et al., 2009), a previous publication from our laboratory (Weiser et al., 2017), and Albritton et al. (2014). Reads were aligned to the genome using HISAT2 (Kim et al., 2015) with the –dta flag. The resulting SAM alignments were converted to BAM format using SAMtools (Li et al., 2009)and provided as input into StringTie2 version 2.0 (Kovaka et al., 2019). StringTie2 was run with the WormBase WS265 GFF3 annotation file provided to guide assembly.

2.6.8 3′-UTR calling

To identify putative 3′UTRs, reads were first grouped by their putative stop codons and any splice junctions that occurred downstream from that stop codon. For each read in each of these groups, the 3′-most base in their alignment was extracted. These end positions were then used to generate a Gaussian kernel density estimate (using the Python package Seaborn, version 0.9.0 kdeplot function with a specified kernel width of 10). Local maxima in this kernel density estimate were identified and reported as a putative 3′-UTR cleavage site if there were at least three read end positions within 10 bp of that local maxima. Reads were assigned to a given 3′UTR if that UTR's putative cleavage site was the closest UTR cleavage site to the end position of the read and if the end position of the read and the putative cleavage site were within 10 bp of each other.

2.6.9 Poly(A) tail length estimation

Poly(A) tail lengths were estimated from raw signal for each read using the poly(A) estimation function of the program nanopolish (version 0.10.2)

(Workman et al., 2019). Poly(A) tail length estimates were only considered if the QC tag reported by nanopolish was PASS. Poly(A) tail length estimates were grouped by gene and isoform using the gene and isoform assignments for each read derived from comparison of genomic alignments with the splice junctions in the WormBase WS265 GFF3 reference.

2.6.10 Calculating coverage for the metagene plot

To generate the metagene plot displayed in Figure 2.1B, we calculated coverage across every gene (as defined by the ce11 WS245 WormBase .gtf annotation file converted to BED format) using the BEDTools coverage function (Quinlan and Hall, 2010; Lee et al., 2018). We then summed these coverage values together and normalized the resulting values by dividing each value by the sum of all the coverage values. Gene sizes were scaled such that the size of the gene body and the UTRs were always the same.

2.6.11 Determining full-length support from WormBase annotations

A WormBase splice isoform was said to have full-length support if every one of its introns in the WS265 annotation GFF3 was annotated to have support from the same EST or the same cDNA (Lee et al., 2018; WormBase web site, 2018). This restricted our analysis to only consider isoforms that were annotated as having introns and excluded single exon genes and genes without introns annotated in the GFF3 annotation file (which includes all noncoding RNAs). To account for this, when comparing the number of genes and isoforms we support to the number of genes and isoforms with full-length

support in WormBase, we only considered splice isoforms from our data set that corresponded to an isoform from the restricted WormBase isoform set. Annotated isoforms that lack support from full-length sequencing may still represent bona fide full-length transcripts whose annotation was derived with the aid of some degree of inference. However, without such empirical sequencing evidence, we cannot be completely confident in calling it a validated full-length transcript (see Supplemental Material for more details).

2.6.12 Predicting coding potential with CPAT

We utilized CPAT (Coding-Potential Assessment Tool) to determine the number of splice isoforms we identify that are predicted to be coding, as well as to filter for reads from isoforms predicted to be coding in Figure 2.4B (Wang et al., 2013). To train this algorithm on the C. elegans transcriptome, we utilized three files from the WS265 WormBase annotation ftp site (Lee et al., 2018; WormBase web site, 2018), the FASTA file describing CDS transcripts, the FASTA file describing mRNA transcripts, and the FASTA file describing ncRNA transcripts. We first converted all Us in the ncRNA FASTAs to Ts using sed 's/U/T/g', and then used the ncRNA FASTA and the CDS FASTA in the CPAT script make_hexamer_tab.py to generate a file of hexamer counts in noncoding and coding RNA in C elegans. We then ran the CPAT script make_logitModel.py using the mRNA FASTA file, the ncRNA FASTA file, and the hexamer count file generated by make_hexamer_tab.py. We used the resulting model as into cpat.py, along with the extracted sequences from each of our splice isoforms, to generate a coding potential prediction for each splice isoform we identify. We used a threshold of 0.5 as our cutoff between

noncoding and coding isoforms.

2.6.13 3′-UTR comparisons

We compared our 3′UTRs to the 3′UTRs identified in Jan et al. (Jan et al., 2011) and Mangone et al. (Mangone et al., 2010) using a custom script, compareUTRdatasets.py (available on the GitHub for this project `https://github.com/NatPRoach/c_elegans_dRNAseq_analysis` and as Supplemental Code), that required putative stop codons match identically but allowed for a 10-bp tolerance in putative 3′UTR end positions (Mangone et al., 2010; Jan et al., 2011). Since previous studies examining 3′UTRs would be unable to identify splicing structure within the 3′UTR, we considered only the chromosome, start, stop, and strand of our 3′UTRs when comparing the number and overlap of 3′UTRs in our data set with these previous data sets. Collapsing the data in this way very slightly reduces the number of unique 3′UTRs in our data set, hence the slight discrepancy between the number of 3′UTRs accounted for in Figure 2.3, A and B, and the number of 3′UTRs reported in Supplemental Table S2.3. We identified novel 3′UTRs in a similar manner but also added consideration of WormBase annotated 3′UTRs.

2.6.14 Calling PAS sites

We identified PAS sites in a method similar to that used by Mangone et al., in which we searched the 60 nt upstream of the putative cleavage site for putative PAS hexamers (Mangone et al., 2010). Rather than recalculating the frequency of putative PAS hexamers upstream of our putative cleavage sites, we used the PAS hexamers specified in Supplemental Table 5 of Mangone

et al. (2010) and searched for these hexamers in the order they appear in that table. Once a putative PAS site was identified, the UTR was assigned that PAS hexamer. If the 3'UTR had none of the hexamers present in the table in its upstream sequence, the UTR was said to have no PAS. Plotting PAS nucleotide distributions To plot the nucleotide distribution around a given type of PAS site, we first sorted sequences by their PAS type. For canonical and alternative PAS sites, nucleotide distributions were anchored such that the PAS site began at 19 nt. The percentage of use of each base at each position in a window around the PAS site was then calculated. For UTRs with no PAS identified, the nucleotide distribution was calculated such that the putative cleavage site was at position 0.

2.7 Data access

All raw and processed sequencing data generated in this study have been submitted to the European Nucleotide Archive (ENA; https://www.ebi.ac.uk/ena) under accession number PRJEB31791. The code required to replicate the analyses performed in this paper is available as Supplemental Code, as well as on GitHub at https://github.com/NatPRoach/c_elegans_dRNAseq_analysis.

2.8 Competing interest statement

Nathan P. Roach, Norah Sadowski, and Winston Timp were reimbursed for conference fees, travel, and accommodation to speak at events organized by Oxford Nanopore Technologies (ONT). Winston Timp has two patents

licensed to ONT (8,748,091 and 8,394,584).

2.9 Acknowledgments

This work was supported by a grant from the National Institutes of Health to John K. Kim (NIH R01GM129301) and to Winston Timp (NIH R01HG010538) and by a Johns Hopkins Discovery Award (Office of the Provost) to Winston Timp, James Taylor, and John K. Kim. Nathan P.Roach was partly supported by a training grant awarded to the Johns Hopkins Cell, Molecular, Developmental Biology and Biophysics program (NIH T32GM007231). We thank Mindy Clark for the *C. elegans* life cycle diagram in Figure 2.1A. We thank Mallory Freeberg and Rachael Workman for initial sequencing and computational analyses comparing nanopore-based cDNA and dRNA sequencing that led us to utilize dRNA-seq in this study. We thank Angela Brooks and Allison Tang for helpful discussions on characterizing full-length transcripts.

2.10 References

Albritton SE, Kranz AL, Rao P, Kramer M, Dieterich C, and Ercan S. 2014. Sex-biased gene expression and evolution of the x chromosome in nematodes. *Genetics* **197**: 865–883.

Andreassi C and Riccio A. 2009. To localize or not to localize: mRNA fate is in 3'UTR ends. *Trends Cell Biol.* **19**: 465–474.

Bayega A, Oikonomopoulos S, Zorbas E, Wang YC, Gregoriou ME, Tsoumani KT, Mathiopoulos KD, and Ragoussis J. 2018. Transcriptome landscape of the developing olive fruit fly embryo delineated by oxford nanopore long-read RNA-Seq.

Blazie SM, Geissel HC, Wilky H, Joshi R, Newbern J, and Mangone M. 2017. Alternative polyadenylation directs Tissue-Specific miRNA targeting in caenorhabditis elegans somatic tissues. *Genetics* **206**: 757–774.

Broverman SA and Meneely PM. 1994. Meiotic mutants that cause a polar decrease in recombination on the X chromosome in caenorhabditis elegans. *Genetics* **136**: 119–127.

Byrne A, Beaudin AE, Olsen HE, Jain M, Cole C, Palmer T, DuBois RM, Forsberg EC, Akeson M, and Vollmers C. 2017. Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. *Nat. Commun.* **8**: 16027.

Cai Y, Yu X, Hu S, and Yu J. 2009. A brief review on the mechanisms of miRNA regulation. *Genomics Proteomics Bioinformatics* **7**: 147–154.

Chang H, Lim J, Ha M, and Kim VN. 2014. TAIL-seq: genome-wide determination of poly(a) tail length and 3′ end modifications. *Mol. Cell* **53**: 1044–1052.

Chen RAJ, R A -, Down TA, Stempor P, Chen QB, Egelhofer TA, Hillier LW, Jeffers TE, and Ahringer J. 2013. The landscape of RNA polymerase II transcription initiation in c. elegans reveals promoter and enhancer architectures.

Corsi AK, Wightman B, and Chalfie M. 2015. A transparent window into biology: A primer on caenorhabditis elegans. *Genetics* **200**: 387–407.

De Coster W, D'Hert S, Schultz DT, Cruts M, and Van Broeckhoven C. 2018. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* **34**: 2666–2669.

Diag A, Schilling M, Klironomos F, Ayoub S, and Rajewsky N. 2018. Spatiotemporal m(i)RNA architecture and 3'UTR regulation in the c. elegans germline.

Garalde DR, Snell EA, Jachimowicz D, Sipos B, Lloyd JH, Bruce M, Pantic N, Admassu T, James P, Warland A, et al.. 2018. Highly parallel direct RNA sequencing on an array of nanopores. *Nat. Methods* **15**: 201–206.

Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, Liu T, Yip KY, Robilotto R, Rechtsteiner A, Ikegami K, et al.. 2010. Integrative analysis of the caenorhabditis elegans genome by the modENCODE project. *Science* **330**: 1775–1787.

Hillier LW, Coulson A, Murray JI, Bao Z, Sulston JE, and Waterston RH. 2005. Genomics in c. elegans: so many genes, such a little worm. *Genome Res.* **15**: 1651–1660.

Hillier LW, Reinke V, Green P, Hirst M, Marra MA, and Waterston RH. 2009. Massively parallel sequencing of the polyadenylated transcriptome of c. elegans. *Genome Res.* **19**: 657–666.

Hodgkin J, Horvitz HR, and Brenner S. 1979. Nondisjunction mutants of the nematode CAENORHABDITIS ELEGANS. *Genetics* **91**: 67–94.

Jan CH, Friedman RC, Ruby JG, and Bartel DP. 2011. Formation, regulation and evolution of caenorhabditis elegans 3'UTRs. *Nature* **469**: 97–101.

Jenjaroenpun P, Wongsurawat T, Pereira R, Patumcharoenpol P, Ussery DW, Nielsen J, and Nookaew I. 2018. Complete genomic and transcriptional landscape analysis using third-generation sequencing: a case study of saccharomyces cerevisiae CEN.PK113-7D. *Nucleic Acids Res.* **46**: e38.

Kadobianskyi M, Schulze L, Schuelke M, and Judkewitz B. 2019. Hybrid genome assembly and annotation of danionella translucida, a transparent fish with the smallest known vertebrate brain.

Kim D, Langmead B, and Salzberg SL. 2015. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**: 357–360.

Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, and Pertea M. 2019. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* **20**: 278.

Kuersten S and Goodwin EB. 2003. The power of the 3′ UTR: translational control and development. *Nat. Rev. Genet.* **4**: 626–637.

Lamesch P, Milstein S, Hao T, Rosenberg J, Li N, Sequerra R, Bosak S, Doucette-Stamm L, Vandenhaute J, Hill DE, et al.. 2004. C. elegans ORFeome version 3.1: increasing the coverage of ORFeome resources with improved gene predictions. *Genome Res.* **14**: 2064–2069.

Lee RC, Feinbaum RL, and Ambros V. 1993. The c. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* **75**: 843–854.

Lee RYN, Howe KL, Harris TW, Arnaboldi V, Cain S, Chan J, Chen WJ, Davis P, Gao S, Grove C, et al.. 2018. WormBase 2017: molting into a new stage. *Nucleic Acids Res.* **46**: D869–D874.

Legnini I, Alles J, Karaiskos N, Ayoub S, and Rajewsky N. 2019. Full-length mRNA sequencing reveals principles of poly (a) tail length control. *bioRxiv* .

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**: 3094–3100.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, and 1000 Genome Project Data Processing Subgroup. 2009. The sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

Lim J, Lee M, Son A, Chang H, and Kim VN. 2016. mTAIL-seq reveals dynamic

poly(a) tail regulation in oocyte-to-embryo development. *Genes Dev.* **30**: 1671–1682.

Lima SA, Chipman LB, Nicholson AL, Chen YH, Yee BA, Yeo GW, Coller J, and Pasquinelli AE. 2017. Short poly(a) tails are a conserved feature of highly expressed genes. *Nat. Struct. Mol. Biol.* **24**: 1057–1063.

Mangone M, Manoharan AP, Thierry-Mieg D, Thierry-Mieg J, Han T, Mackowiak SD, Mis E, Zegar C, Gutwein MR, Khivansara V, et al.. 2010. The landscape of c. elegans 3′ UTRs. *Science* **329**: 432–435.

Mayr C and Bartel DP. 2009. Widespread shortening of 3′UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell* **138**: 673–684.

Millonigg S, Minasaki R, Nousch M, and Eckmann CR. 2014. GLD-4-Mediated translational activation regulates the size of the proliferative germ cell pool in the adult c. elegans germ line.

Nousch M, Minasaki R, and Eckmann CR. 2017. Polyadenylation is the key aspect of GLD-2 function in c. elegans. *RNA* **23**: 1180–1187.

Packer JS, Zhu Q, Huynh C, Sivaramakrishnan P, and others. 2019. A lineage-resolved molecular atlas of c. elegans embryogenesis at single cell resolution. *bioRxiv* .

Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, and Salzberg SL. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**: 290–295.

Phillips CM, Wong C, Bhalla N, Carlton PM, Weiser P, Meneely PM, and Dernburg AF. 2005. HIM-8 binds to the X chromosome pairing center and mediates chromosome-specific meiotic synapsis. *Cell* **123**: 1051–1063.

Quinlan AR. 2014. BEDTools: the swiss-army tool for genome feature analysis. *Curr. Protoc. Bioinformatics* **47**: 11–12.

Quinlan AR and Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842.

Reboul J, Vaglio P, Tzellas N, Thierry-Mieg N, Moore T, Jackson C, Shin-i T, Kohara Y, Thierry-Mieg D, Thierry-Mieg J, et al.. 2001. Open-reading-frame sequence tags (OSTs) support the existence of at least 17,300 genes in c. elegans. *Nat. Genet.* **27**: 332–336.

Saito TL, Hashimoto SI, Gu SG, Morton JJ, Stadler M, Blumenthal T, Fire A, and Morishita S. 2013. The transcription start site landscape of c. elegans. *Genome Res.* **23**: 1348–1361.

Schmid M, Küchler B, and Eckmann CR. 2009. Two conserved regulatory cytoplasmic poly(a) polymerases, GLD-4 and GLD-2, regulate meiotic progression in c. elegans. *Genes Dev.* **23**: 824–836.

Sessegolo C, Cruaud C, Da Silva C, Dubarry M, Derrien T, Lacroix V, and Aury JM. 2019. Transcriptome profiling of mouse samples using nanopore sequencing of cDNA and RNA molecules.

Soneson C, Yao Y, Bratus-Neuenschwander A, Patrignani A, Robinson MD,

and Hussain S. 2019. A comprehensive examination of nanopore native RNA sequencing for characterization of complex transcriptomes.

Spieth J, Genome Sequencing Center, Washington University School of Medicine, Louis S, and Usa MO. 2014. Overview of gene structure in c. elegans.

Stiernagle T. 2006. Maintenance of c. elegans. WormBook. the c. elegans research community. *WormBook* .

Subtelny AO, Eichhorn SW, Chen GR, Sive H, and Bartel DP. 2014. Poly(A)-tail profiling reveals an embryonic switch in translational control.

Suh N, Jedamzik B, Eckmann CR, Wickens M, and Kimble J. 2006. The GLD-2 poly (a) polymerase activates gld-1 mRNA in the caenorhabditis elegans germ line. *Proceedings of the National Academy of Sciences* **103**: 15108–15112.

Sulston JE, Schierenberg E, White JG, and Thomson JN. 1983. The embryonic cell lineage of the nematode caenorhabditis elegans. *Dev. Biol.* **100**: 64–119.

Szostak E and Gebauer F. 2013. Translational control by 3'-UTR-binding proteins. *Brief. Funct. Genomics* **12**: 58–65.

Tang AD, Soulette CM, van Baren MJ, Hart K, and others. 2018. Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *bioRxiv* .

The C elegans Sequencing Consortium. 1998. Genome sequence of the nematode c. elegans: A platform for investigating biology.

Tourasse NJ, Millet JRM, and Dupuy D. 2017. Quantitative RNA-seq meta-analysis of alternative exon usage in c. elegans. *Genome Res.* **27**: 2120–2128.

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, and Pachter L. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and cufflinks. *Nat. Protoc.* **7**: 562–578.

Volden R, Palmer T, Byrne A, Cole C, Schmitz RJ, Green RE, and Vollmers C. 2018. Improving nanopore read accuracy with the R2C2 method enables the sequencing of highly multiplexed full-length single-cell cDNA.

Walhout AJM, Temple GF, Brasch MA, Hartley JL, Lorson MA, van den Heuvel S, and Vidal M. 2000. [34] GATEWAY recombinational cloning: Application to the cloning of large numbers of open reading frames or ORFeomes. In *Methods in Enzymology* (eds. J Thorner, SD Emr, and JN Abelson), volume 328, pp. 575–IN7. Academic Press.

Wang L, Park HJ, Dasari S, Wang S, Kocher JP, and Li W. 2013. CPAT: Coding-Potential assessment tool using an alignment-free logistic regression model. *Nucleic Acids Res.* **41**: e74.

Weiser NE, Yang DX, Feng S, Kalinava N, Brown KC, Khanikar J, Freeberg MA, Snyder MJ, Csankovszki G, Chan RC, et al.. 2017. MORC-1 integrates nuclear RNAi and transgenerational chromatin architecture to promote germline immortality. *Dev. Cell* **41**: 408–423.e7.

West SM, Mecenas D, Gutwein M, Aristizábal-Corrales D, Piano F, and Gunsalus KC. 2018. Developmental dynamics of gene expression and alternative polyadenylation in the caenorhabditis elegans germline. *Genome Biol.* **19**: 8.

Wightman B, Bürglin TR, Gatto J, Arasu P, and Ruvkun G. 1991. Negative regulatory sequences in the lin-14 3′-untranslated region are necessary to generate a temporal switch during caenorhabditis elegans development. *Genes Dev.* **5**: 1813–1824.

Williams GW, Davis PA, Rogers AS, Bieri T, Ozersky P, and Spieth J. 2011. Methods and strategies for gene structure curation in WormBase. *Database* **2011**: baq039.

Wilson RK. 1999. How the worm was won. the c. elegans genome sequencing project. *Trends Genet.* **15**: 51–58.

Workman RE, Tang AD, Tang PS, Jain M, Tyson JR, Razaghi R, Zuzarte PC, Gilpatrick T, Payne A, Quick J, et al.. 2019. Nanopore native RNA sequencing of a human poly(a) transcriptome. *Nat. Methods* **16**: 1297–1305.

WormBase web site. 2018. http://wormbase.org, release WS265.

Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Girón CG, et al.. 2018. Ensembl 2018. *Nucleic Acids Res.* **46**: D754–D761.

Zhao HQ, Zhang P, Gao H, He X, Dou Y, Huang AY, Liu XM, Ye AY, Dong MQ, and Wei L. 2015. Profiling the RNA editomes of wild-type c. elegans and ADAR mutants. *Genome Res.* **25**: 66–75.

2.11 Supplemental Material

2.11.1 Supplemental Discussion

2.11.1.1 Regarding identification of trans splice sites with dRNA-seq

Splicing in trans is a common RNA processing event in *C. elegans*, and identifying trans splice sites would in theory be one way of identifying full-length RNA transcripts. Though it is possible to identify putative splice leader sequences at the 5′ end of transcripts with direct RNA sequencing, it is an extremely challenging and error prone task for the technology. To start, the last 10 - 15 bases at the 5′ end of each transcript are not read by the sequencer. This means, in the best case scenario, only the 12 of the 3′ most nts of the 22nt splice leader will be registered. In addition, the error rate of nanopore sequencing ranges between 10 - 15%, and as stated in the main text, in our sequencing experiments averaged at around 14%. This means that, on average, there will be at least one error in those 12 bases. These errors are predominantly insertions and deletions, error types that are essentially not considered in the logic of motif finding and motif matching approaches. All of this contributes to the difficulty of accurately determining if transcripts contain 5′ SL sequences. Thus, although one could likely identify true positive transcripts trans spliced to SL sequences, the lack of a matching motif does not necessarily imply the lack of trans splicing of that isoform, as the truncated 5′ ends and the high error rate ensures that many genuinely trans spliced transcripts will not be identified by motif searching approaches. As such, we opted not to characterize trans splicing in this manuscript.

2.11.1.2 Regarding the "full-length" status of transcripts in the annotation

It should be noted that most existing annotation isoforms are likely "full-length" in that they likely represent full-length transcripts that can be expressed by the organism. However, most of these isoforms are assembled using some amount of inference because the sequencing reads used to support those isoforms are not full-length. As such, though annotation approaches have inferred that these transcripts could be expressed, for many of the annotation isoforms there is no definitive evidence that the full-length isoform is expressed. The best short-read transcriptome assemblers can do to provide support for individual isoforms longer than their read length is infer (using imperfect algorithms) which exons are spliced together in the same isoforms. This is a fundamental problem with short read transcriptome assembly and annotation that can only be addressed using long reads as done in this manuscript.

2.11.2 Supplemental Figures

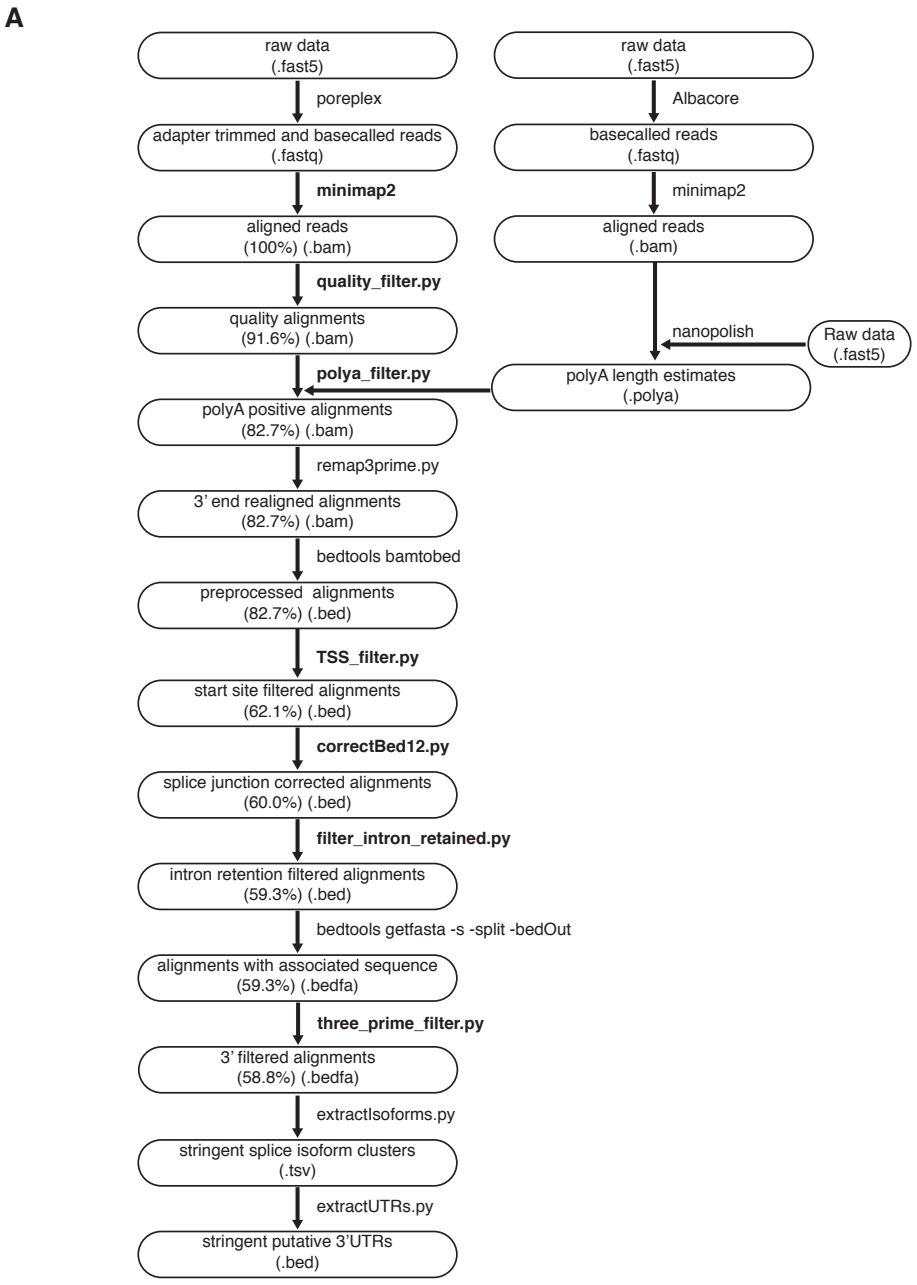
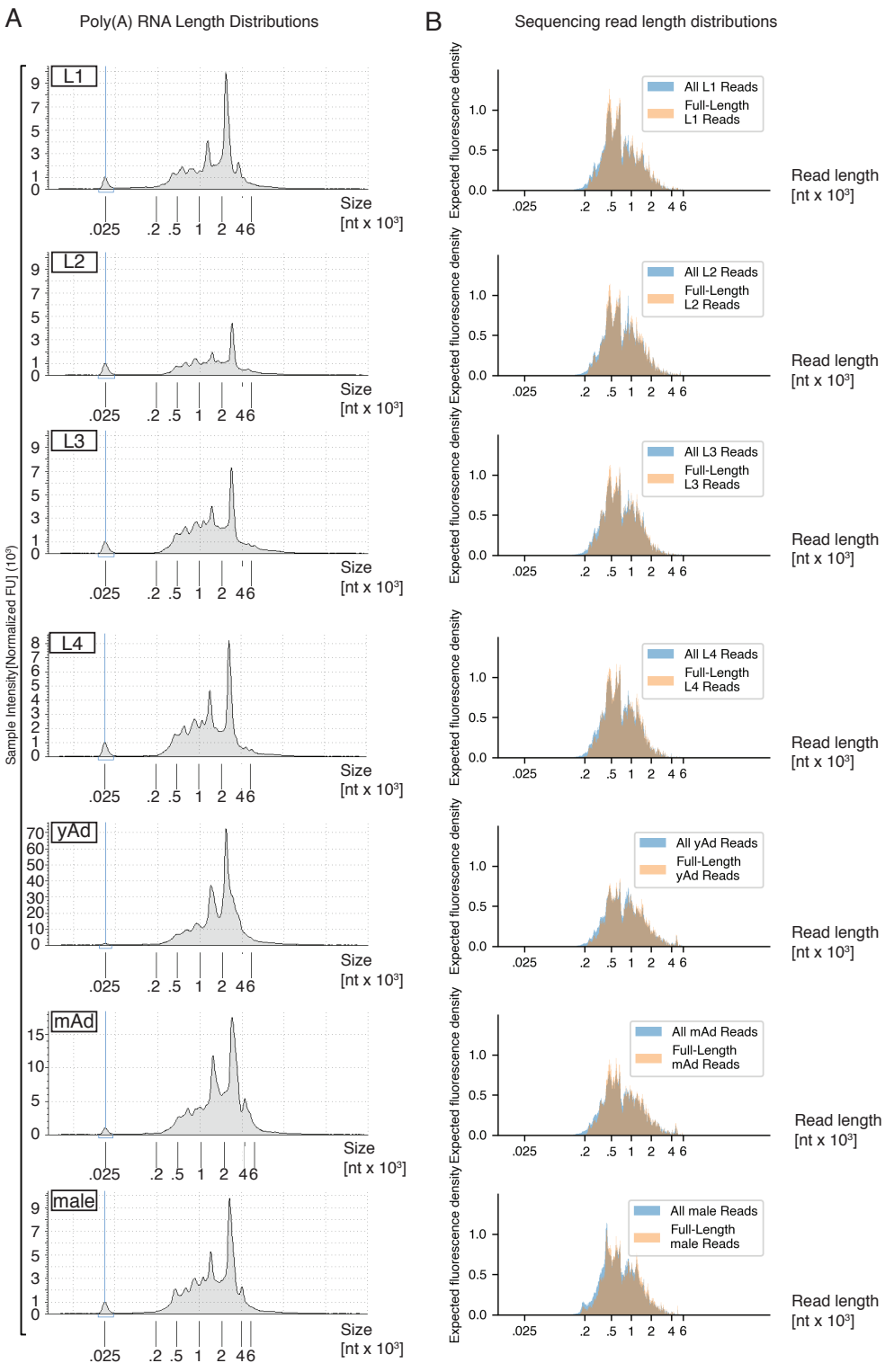


Figure S2.1: (A) Flowchart of analysis pipeline and read filtering used in this study. Percentages indicate the number of aligned reads retained up to that filtering step. File types after each step included in parenthesis.

Figure S2.2 ***(following page)*****: (A)** TapeStation traces showing length distribution of poly(A) selected RNA from each of the developmental stages sequenced. **(B)** Expected fluorescence distribution of reads obtained from dRNAseq of each developmental stage before and after filtering steps were applied.


A
Poly(A) RNA Length Distributions
B
Sequencing read length distributions
Sample Intensity[Normalized FU] (10³)
L1
L2
L3
L4
yAd
mAd
male
Size [nt x 10³]
.025 .2 .5 1 2 4 6
Expected fluorescence density
All L1 Reads
Full-Length L1 Reads
All L2 Reads
Full-Length L2 Reads
All L3 Reads
Full-Length L3 Reads
All L4 Reads
Full-Length L4 Reads
All yAd Reads
Full-Length yAd Reads
All mAd Reads
Full-Length mAd Reads
All male Reads
Full-Length male Reads
Read length [nt x 10³]

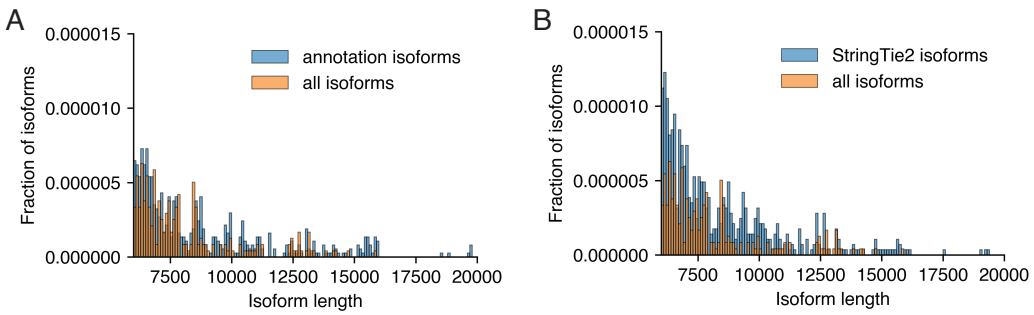


Figure S2.3: Histograms comparing isoform length densities at high lengths **(A)** Comparison of length distributions of isoforms present in the WormBase WS265 annotation, and splice isoforms identified by this study displayed as a density plot **(B)** As in A, comparison of length distribution of isoforms assembled by StringTie2 using Illumina based RNA-seq from across *C elegans* development, and splice isoforms identified by this study.

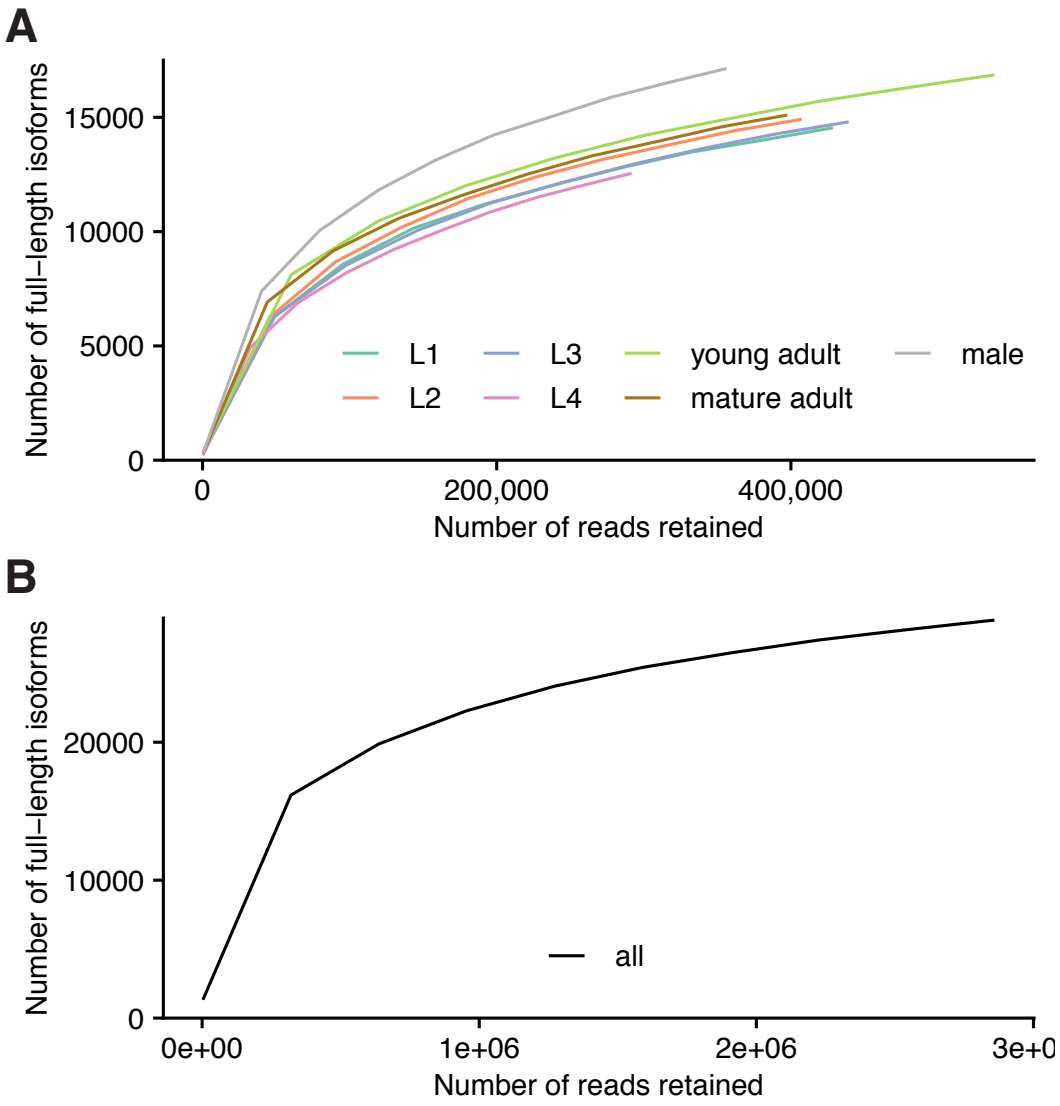


Figure S2.4: (A) Saturation plot showing the number of full-length isoforms with support from one or more reads versus the number of reads considered, separated by stage. **(B)** As in (A), but with all stages combined.

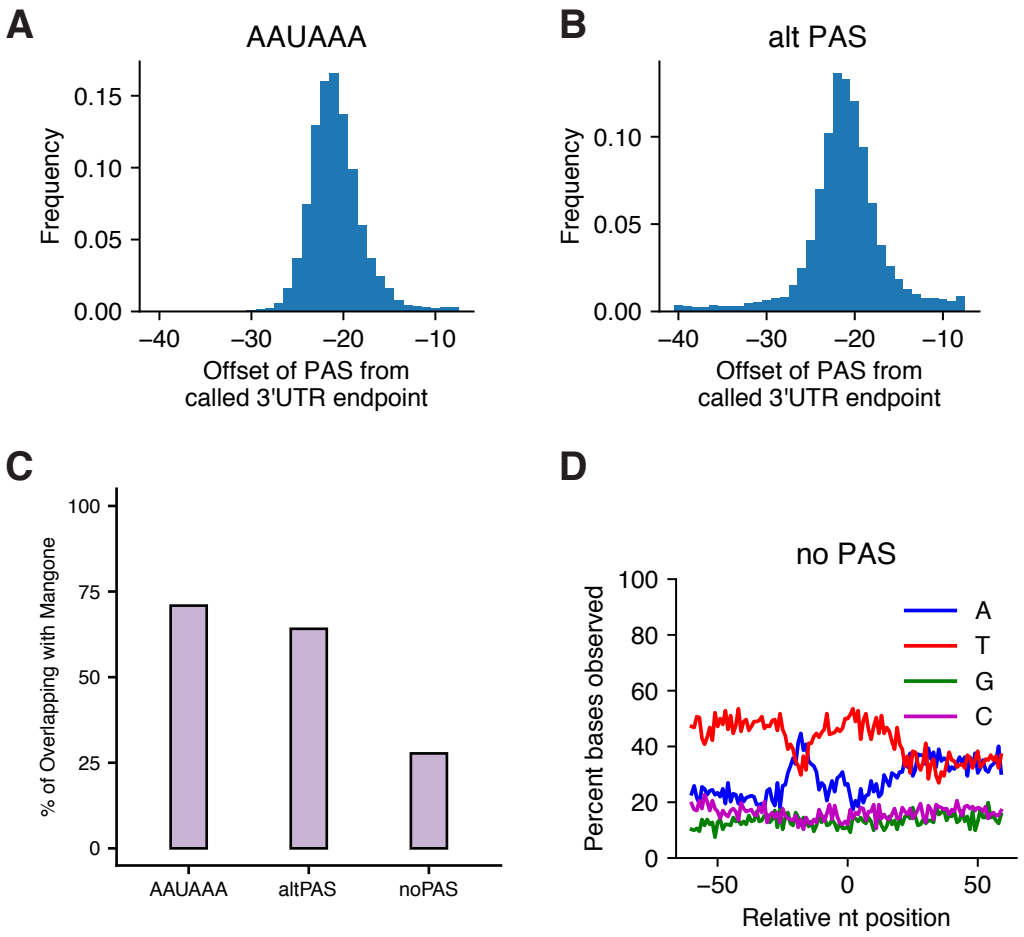


Figure S2.5: Evidence supporting the validity of our identified 3′UTRs. Offsets of identified PAS sites from the putative cleavage site for canonical **(A)** and non-canonical **(B)** PAS sites. **(C)** Percent of UTRs with specified PAS site type that overlap with a Mangone et al. 3′UTR. **(D)** Nucleotide distribution in a window around putative cleavage sites for 3′UTRs that overlap with a Mangone 3′UTR and do not have a PAS site identified. This distribution is different than the published distribution of no PAS Mangone 3′UTRs in general (Mangone et al. 2010)

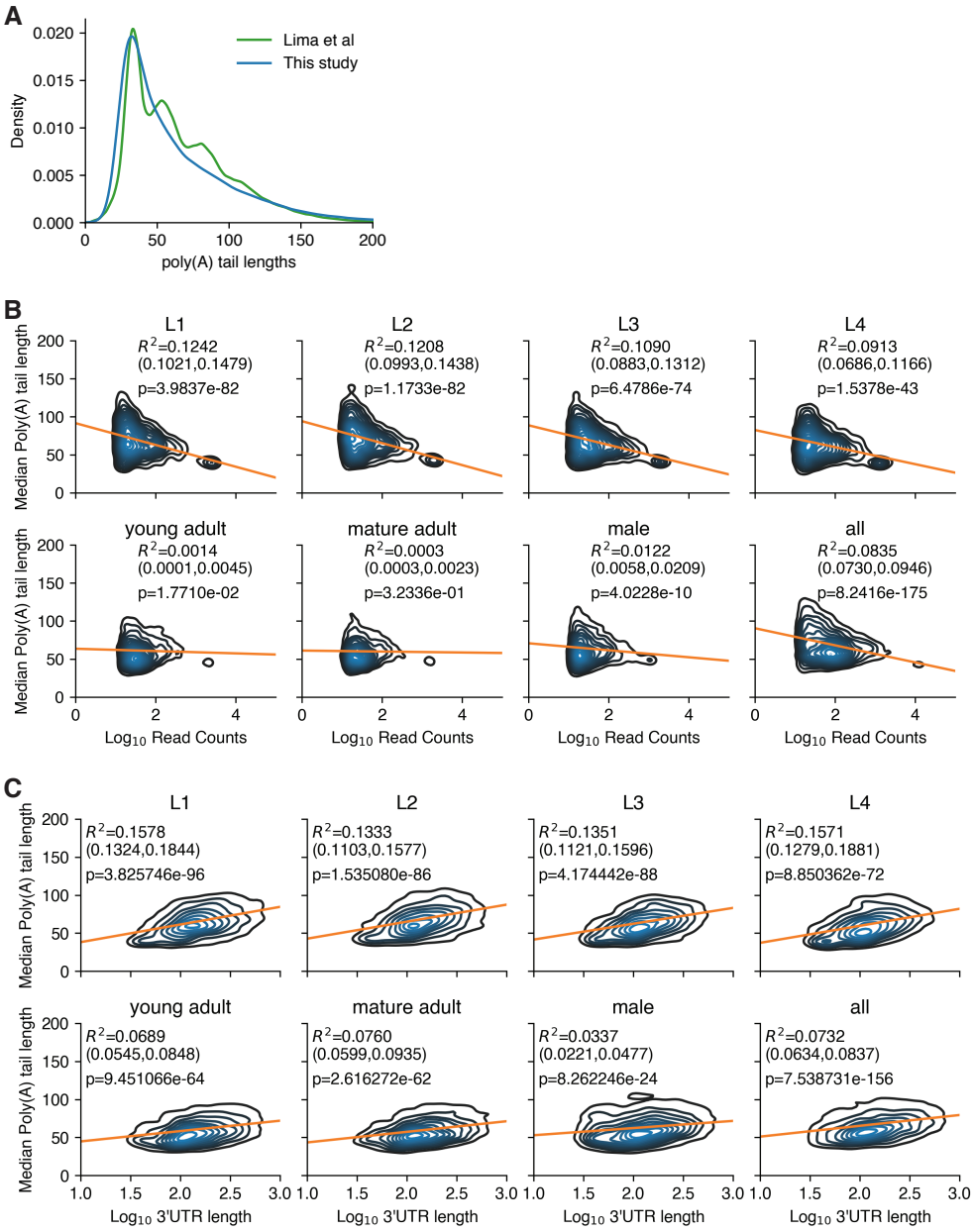


Figure S2.6: (A) Comparison of poly(A) tail length distributions between reads from our L4 stage dataset and Lima et al. (Lima et al. 2017). Density plots including linear regressions (orange line) of median poly(A) tail length versus expression level **(B)** or 3'UTR length **(C)**, separated by stage. Parenthesis indicate 95% confidence intervals for R2 values. P-values calculated on Pearson correlation coefficients.

2.11.3 Supplemental Tables

Table S2.1: Quality control metrics for dRNA-seq samples, provided as a supplemental .xlsx file to this document

Table S2.2: Filtering statistics for dRNA-seq samples, provided as a supplemental .xlsx file to this document

Table S2.3: Isoform statistics for dRNA-seq samples, provided as a supplemental .xlsx file to this document

Stage	Genes with significant correlation between 3′ UTR and splice isoform
L1	2
L2	1
L3	5
L4	2
young adult	9
mature adult	9
male	10

Table S2.4: Number of genes with correlations between 3′UTR and splice isoforms for each sequenced stage

Table S2.5: List of splice isoforms predicted to be non-coding, provided as a supplemental .csv file to this document

Table S2.6: Accession numbers and citations for Illumina data used in this study, provided as a supplemental .xlsx file to this document

Chapter 3

CONDUIT: A short and long read hybrid reference-free transcriptome assembler

3.1 Citation

Roach NP, Fan Y, Starostik MR, Timp W, Schatz MC, Kim JK, and Taylor J. 2020. CONDUIT: A short and long read hybrid reference-free transcriptome assembler. *In submission* .

3.2 Author Contributions

N.P.R. designed and implemented CONDUIT in consultation with J.T., M.C.S. and J.K.K., and wrote the manuscript in consultation with M.R.S, J.K.K and M.C.S.. N.P.R., M.R.S, M.C.S. and J.K.K. reviewed the manuscript. Y.F. collected the *Candida nivariensis* data and assembled the *nivariensis* genome in consultation with W.T.

3.3 Abstract

Reconstructing the transcriptome from RNA sequencing reads is a challenging problem, especially when no high quality reference genome is available. Short read sequencing technologies, though accurate, cannot reliably reconstruct full-length transcripts due to the highly complex nature of the transcriptome with large gene families, widespread alternative splicing, and highly variable expression and coverage per transcript. Meanwhile, single molecule long read sequencing, though capable of producing much longer reads, is highly error prone, and attempts to reconstruct full-length transcripts with these technologies typically contain errors.

Here, we present a novel open-source transcriptome assembler, CONDUIT (https://github.com/NatPRoach/conduit), which uses single molecule long read RNA sequencing to generate scaffolded splice graphs independent of a reference genome. It then pseudomaps short-read RNA sequencing reads to isoforms extracted from the scaffolded splice graph, polishes these splice graphs using both short and long read data, and outputs consensus isoforms extracted from these splice graphs. We show that CONDUIT produces highly accurate consensus isoforms, completely independent of a reference genome in several model systems and in a novel pathogenic yeast system. The isoforms assembled from sequencing data from the *C. elegans* L4 stage had a 99.9% median percent reference identity, of which 64.1% had perfect intron chain level correspondence to an isoform annotated in the ce11 transcriptome annotation, and 67.1% of in silico predicted protein products perfectly corresponded to a *C. elegans* protein annotated in Uniprot. Moreover, CONDUIT was more accurate

at predicting proteins in a reference-free manner than existing reference-free transcriptome assemblers RATTLE, Trinity, and rnaSPAdes, and the long-read transcription-aware RNA-seq correction tool TALC, and in some scenarios outcompeted the *ab initio* transcriptome assembly method StringTie2. For example, in the pathogenic yeast *Candida nivariensis*, an organism without a high quality reference genome, 91% of CONDUIT protein predictions and 77.1% of StringTie2 predictions matched with an annotated protein from the closely related species *Candida glabrata* in a BLASTP search.

3.4 Introduction

RNA sequencing (RNA-seq) is an indispensable tool for the purposes of annotating the complete set of RNA products produced by an organism (i.e. the transcriptome) (Stark et al., 2019). Standard RNA-seq protocols are broadly useful for the assembly of transcriptomes, and for quantifying RNA products thereby allowing for differential expression analysis between sequencing samples, while more specialized RNA-seq protocols have been used to annotate 3-prime untranslated region structures, transcription start sites, poly(A) tail lengths, and much more (Mangone et al., 2010; Jan et al., 2011; Saito et al., 2013; Lima et al., 2017).

Transcriptome assemblers broadly fall into one of two classes, *de novo* assemblers and *ab initio* assemblers. *De novo* assembly involves reconstructing the transcriptome independent of a reference genome sequence, and *ab initio* assembly relies on a genome sequence to improve transcriptome assembly and annotation. While *ab initio* methods rely on alignment of RNA-seq reads

to the reference genome in order to define transcript structure in relation to the genome sequence, *de novo* methods need a means of grouping reads that originate from the same gene in order to determine the isoforms expressed at that gene. The approaches used for such grouping differs depending on whether the method integrates short reads, long reads, or both.

Short read RNA sequencing generates sequencing reads typically around 75 - 150 nt long with low error rates. These short reads are incapable of spanning the full length of most RNA transcripts, and therefore construction of transcriptomes through short read RNA-seq relies on some level of inference, and computational methods for reconstructing putative full-length isoforms *post-hoc*. In the case of *de novo* reference-free transcriptome assembly, such as produced by the Trinity transcriptome assembler, this involves the generation and resolution of weighted De Bruijn graphs (in theory one per gene) (Grabherr et al., 2011). However, the repetitive nature of transcriptomes often results in excessive branching and complex cycles in these De Bruijn graphs leading to erroneously reconstructed transcripts. As such, when high quality reference genomes are available *ab initio* methods for transcriptome assembly such as StringTie2 are often preferred (Kovaka et al., 2019).

Long read RNA-seq, such as produced through Oxford Nanopore Technologies (ONT) or Pacific Biosciences (PacBio) sequencing platforms on the other hand have significantly longer read lengths such that these reads are capable of spanning the full length of RNA-transcripts. These long read lengths allow for the annotation of transcript isoforms without the need to computationally reconstruct these isoforms. However, the utility of PacBio sequencing in

transcriptome annotation and assembly is currently limited by its relatively low throughput, and relatively high cost per base sequenced. In addition, both PacBio and ONT based long read RNA-seq has a higher error rate than short read RNA-seq (10 - 15 % for raw PacBio data and anywhere from 5 - 15% for ONT data depending on the pore chemistry and basecalling model used) (Weirather et al., 2017; Rang et al., 2018; Wick et al., 2019; Laver et al., 2015). It has been shown, however, that consensus sequences (generated by reading the same sequence multiple times by the sequencer) have substantially lower error rates than raw reads in both PacBio and ONT sequencing (Wenger et al., 2019; Volden et al., 2018).

The high error rates associated with long read RNA-seq introduce unique challenges to transcriptome assembly with these technologies. These errors, enriched in insertions and deletions relative to the reference lead to errors in alignment and therefore errors in the resulting intron chains often used to define isoforms. Various *ab initio* transcriptome assembly methods for long read RNA-seq have been generated, most filtering reported splice junctions based on some form of external validation, relying on a high quality existing transcriptome annotation, paired Illumina RNA-seq data, or both (Tang et al., 2018; Wyman et al., 2019; Kovaka et al., 2019). The RATTLE transcriptome assembler (de la Rubia et al. 2020), in contrast, generates *de novo*, reference-free, transcriptome assemblies using long read PacBio or ONT RNA-seq alone. However, the error rates of these single molecule long read technologies are often too high to reconstruct transcripts perfectly, such that there are no errors in the transcripts.

By leveraging the strengths and mitigating the weaknesses of both long and short read sequencing one should in principle generate results more accurate than can be attained with either technology individually. Several tools have been generated for the purposes of utilizing short read sequences to correct long reads prior to alignment. However, most long read correction tools are not designed for use with RNA-seq reads, with the notable exception of TALC (Transcription Aware Long-read Correction) (Broseus et al., 2020). In addition, it is worth noting that for such long-read correction tools alignment to a reference genome is still required for the purposes of defining transcript structure and the relationship between isoforms. The SPAdes assembly algorithm recently released an update that generates hybrid *de novo* transcriptome assemblies by generating de Bruijn graphs, aligning long reads to these graphs, and using these alignments to determine the paths taken through the graph (Prjibelski et al., 2020). This rnaSPAdes approach is currently the only tool explicitly designed for reference genome independent transcriptome assembly that integrates both short and long read data to leverage the strengths of each technology in tandem.

Here we present CONDUIT (CONsensus Decomposition Utility In Transcriptome-assembly), an open-source transcriptome assembler that uses long read RNA-seq data clustered by their gene of origin and corresponding short read RNA-seq data to generate representative consensus isoforms independent of a reference genome (Figure 3.1 & 3.2). This long and short read hybrid *de novo* transcriptome assembler generates highly accurate isoforms by leveraging the strengths of long reads to define isoform structure,

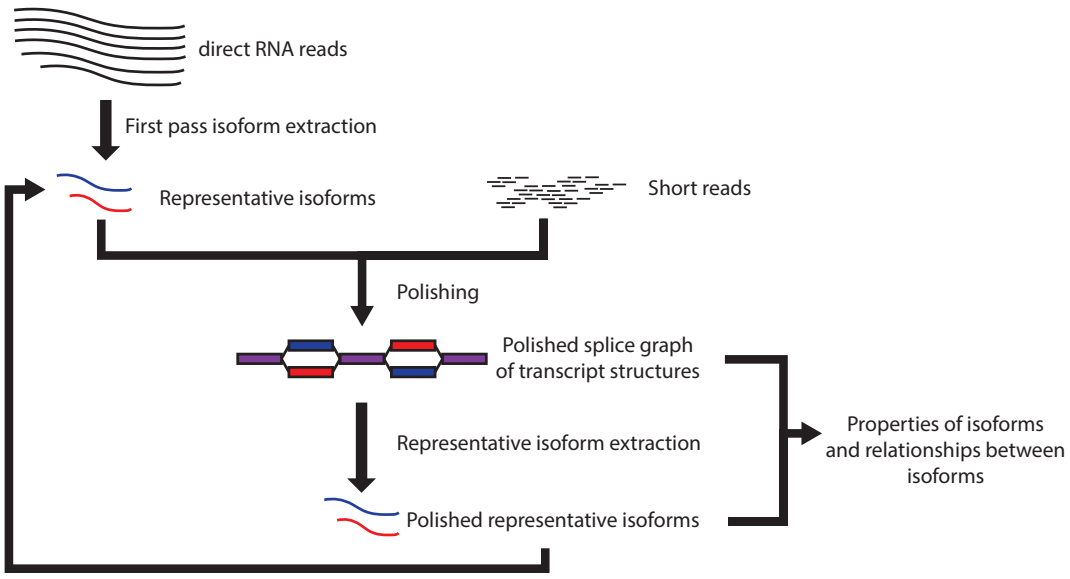


Figure 3.1: Overview of CONDUIT. Long read scaffolds are used to generate putative representative isoforms. The resulting splice graph defining these isoforms and their relationship to one another are then polished with short read RNA-seq data, the representative isoforms are extracted, and this process is repeated for several iterations.

and the strengths of short reads to polish away errors in the long read scaffolds. Once polished representative isoforms are generated properties of these highly accurate isoforms, such as *in silico* predicted protein products, can be determined.

3.5 Results

3.5.1 CONDUIT Algorithm Overview

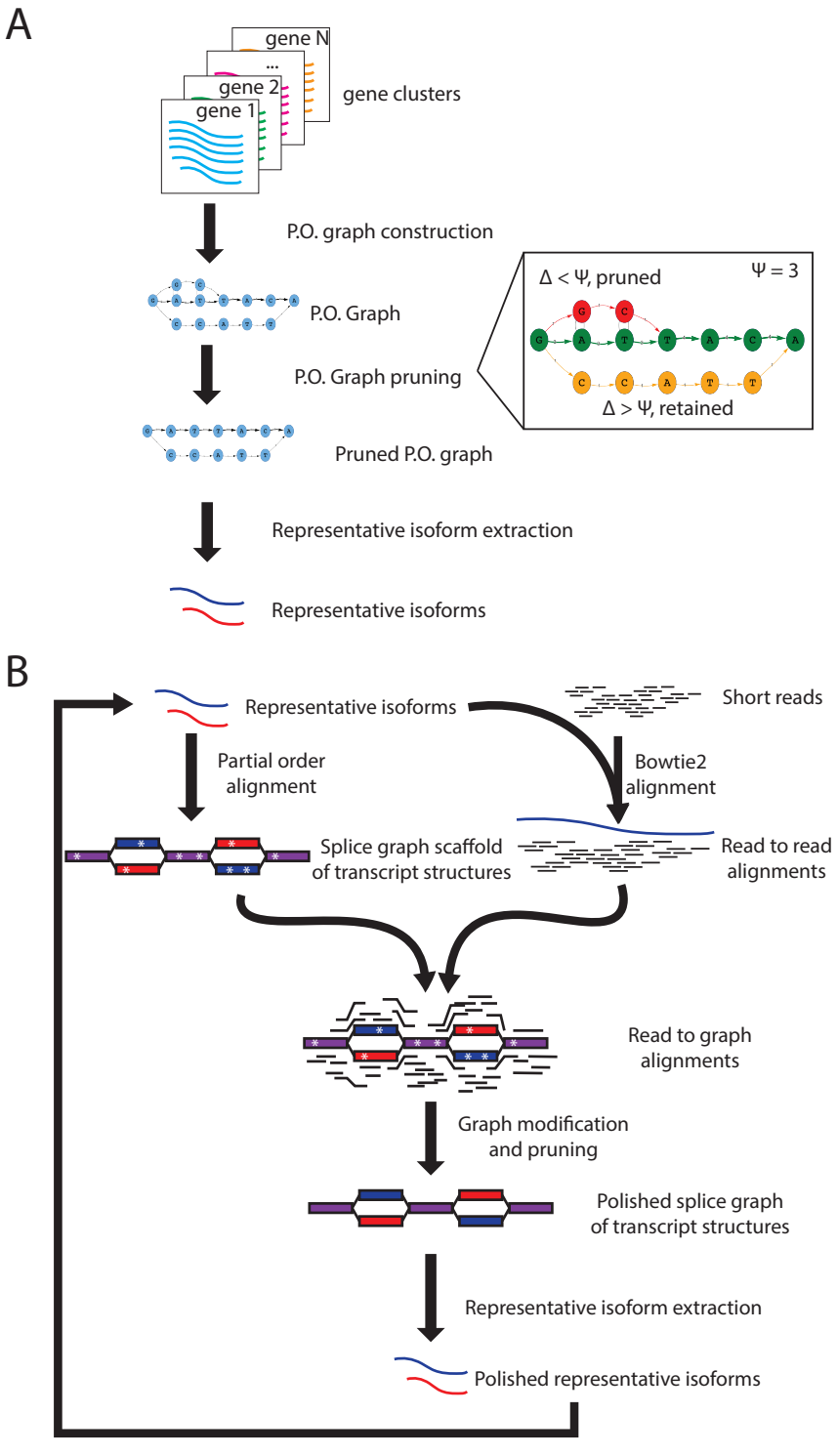
CONDUIT requires as input long read RNA-seq reads clustered at the gene level and corresponding short read RNA sequencing data. Broadly, CONDUIT first extracts putative representative isoforms from the long read RNA-seq clusters, and polishes these isoforms for several iterations using short read

RNA-seq (Figure 3.1).

To accomplish this analysis, CONDUIT first compares the reads in each gene cluster to one another using poaV2 (https://github.com/tanghaibao/bio-pipeline/tree/master/poaV2), an implementation of the partial order alignment algorithm resulting in a Partial Order Graph data 3.2A) (Lee et al., 2002). The partial order graph representation of these alignments is then pruned in a manner such that reads that differ from one another by greater than some number (Ψ) of continuous nucleotides (default 35) are treated as different isoforms, while regions that differ from one another by less than Ψnt are "corrected" to reflect a consensus sequence (see Methods, Section 3.7.5 for detailed description of pruning approach). Once the partial order graph is pruned, representative isoforms are extracted from it in a manner similar to how isoforms are extracted from splice graphs by StringTie (Pertea et al., 2015; Kovaka et al., 2019).

Illumina reads are then aligned to these representative isoforms by Bowtie2, and the partial order graph of the representative isoforms is calculated (Figure 3.2B) (Langmead and Salzberg, 2012). The known path through the graph taken by each representative isoform and the known location of Illumina

Figure 3.2 ***(following page)*****:** Implementation details of CONDUIT. **(A)** Gene clusters are used to generate partial order graphs. These graphs are then pruned to remove putative errors, and representative isoforms are extracted from the graph. **(B)** Representative isoforms are used to construct splice graphs using partial order alignment, and Bowtie2 is used to align short read data to these representative isoforms. These pieces of information are combined to relate Illumina reads to locations in the splice graph. The splice graph is modified to reflect Illumina data, the graph is pruned to remove putative errors, and the representative isoforms from the graph are extracted. This process is then repeated for several iterations.


A
gene N
...
gene 2
gene 1
gene clusters
P.O. graph construction
P.O. Graph
P.O. Graph pruning
Pruned P.O. graph
Δ < Ψ, pruned
Ψ = 3
Δ > Ψ, retained
Representative isoform extraction
Representative isoforms
B
Representative isoforms
Short reads
Partial order alignment
Bowtie2 alignment
Splice graph scaffold of transcript structures
Read to read alignments
Read to graph alignments
Graph modification and pruning
Polished splice graph of transcript structures
Representative isoform extraction
Polished representative isoforms

alignments in each representative isoform is combined to determine the location within the partial order graph supported by the Illumina reads. The graph is modified to reflect this Illumina support, adding new nodes and new edges where the Illumina reads contain insertions, mismatches, and deletions relative to the representative isoforms, and updating the weights of the partial order graph to reflect the Illumina supported edges through the graph. The graph is then pruned and representative isoforms are extracted in the same manner as before. This process of polishing, pruning, and isoform extraction is repeated several times (default: 5) resulting in highly accurate consensus transcript sequences. After these repeated rounds of graph based polishing, a final round of polishing is performed, this time using a linear polishing approach in which each isoform is polished separately. When run in stringent mode (on by default) CONDUIT then filters the isoforms that get reported for only those isoforms supported by short read data in every internal position (see Methods). This iterative polishing and stringent filter substantially improves correspondence of extracted isoforms to the reference genome (Supplemental Figure S3.1, Supplemental Table S3.1)

3.5.2 Performance with Model Organisms

In order to evaluate the efficacy of CONDUIT in reconstructing transcriptomes independently of a reference genome using short and long read RNA-seq data we made use of several existing RNA-seq datasets, summarized in Supplemental Table 2. These datasets were collected in organisms with well annotated genomes and transcriptomes such that we could compare the reconstructed transcriptome with that of the annotation. Overall the datasets with high

quality reference genomes evaluated include: 1-7) *Caenorhabditis elegans* data collected from stages of continuous hermaphroditic development and adult males (Roach et al., 2020; Zhou et al., 2019; Albritton et al., 2014; Niu et al., 2020; Xu et al., 2018; Yang et al., 2019) 8) data from the GM12878 human B-lymphocyte cell line (Workman et al., 2019; Tilgner et al., 2014). In addition to evaluating datasets with high quality reference genomes, we also utilized direct RNA-seq (dRNA-seq) data and Illumina short read RNA-seq from *Candida nivariensis*, a pathogenic yeast with no previously published annotated reference genome (see below).

We focused on four principle metrics for evaluating CONDUIT in samples with a high quality reference genome: 1) runtime 2) percent reference identity upon alignment to the reference genome 3) intron chain level precision / recall relative to the transcriptome annotation and 4) predicted protein sequence level precision / recall relative to the annotation.

We compared CONDUIT's performance with that of five other tools: the Trinity *de novo* transcriptome assembler designed for short read RNA sequencing data, the RATTLE *de novo* transcriptome assembler designed for long read RNA sequencing data, the rnaSPAdes *de novo* hybrid transcriptome assembler, the TALC error correction tool designed for hybrid correction of long reads using short read sequencing, and the StringTie2 *ab initio* transcriptome assembler.

The first metric we evaluated for each *de novo* assembly tool was the runtime from raw data to the final output of each tool on the C. elegans L4 data (Figure 3.3A). Since CONDUIT requires reads partitioned into gene level clusters we

included in the runtime calculations for CONDUIT the runtime of RATTLE gene level clustering. Comparably, we included the runtime of Jellyfish k-mer counting necessary to run TALC in the calculations of TALC's runtime.

As can be seen in Figure 3.3A, all tools complete within 100 to 450 minutes. CONDUIT runs on the *C. elegans* L4 datasets in 7 hours 4 minutes when using 40 threads on a Dell PowerEdge R930 with 4x Intel Xeon E7-4850 v3 @ 2.20 GHz CPUs (56 cores, 112 threads total), and 1,536 Gb of RAM. This runtime is longer relative to existing tools RATTLE, Trinity, TALC, and rnaSPAdes as expected given the relatively long time required to preprocess reads into gene level clusters, and the integration of both short and long read data involving several rounds of short read alignment to long read scaffolds. Notably, TALC ran the fastest, which is expected given that TALC is designed for long read correction not *de novo* transcriptome assembly, and therefore corrects at the individual read level and does not require gene level clustering or partial order alignment of reads corresponding to the same gene or isoform, a computationally expensive step utilized by RATTLE and CONDUIT. The next metric each tool was evaluated on was the percent reference identity of extracted isoforms for the real *C. elegans* L4 dataset (Figure 3.3B). This metric reflects the nucleotide level correspondence to the ce11 reference genome upon alignment of extracted isoforms as calculated by the NanoComp software (De Coster et al., 2018). Note that NanoComp doesn't consider soft-clipped bases in its calculation of percent reference identity, so isoforms with large soft-clips can still be considered perfect matches to the reference genome using this approach. The breakdown of median and average percent reference identities

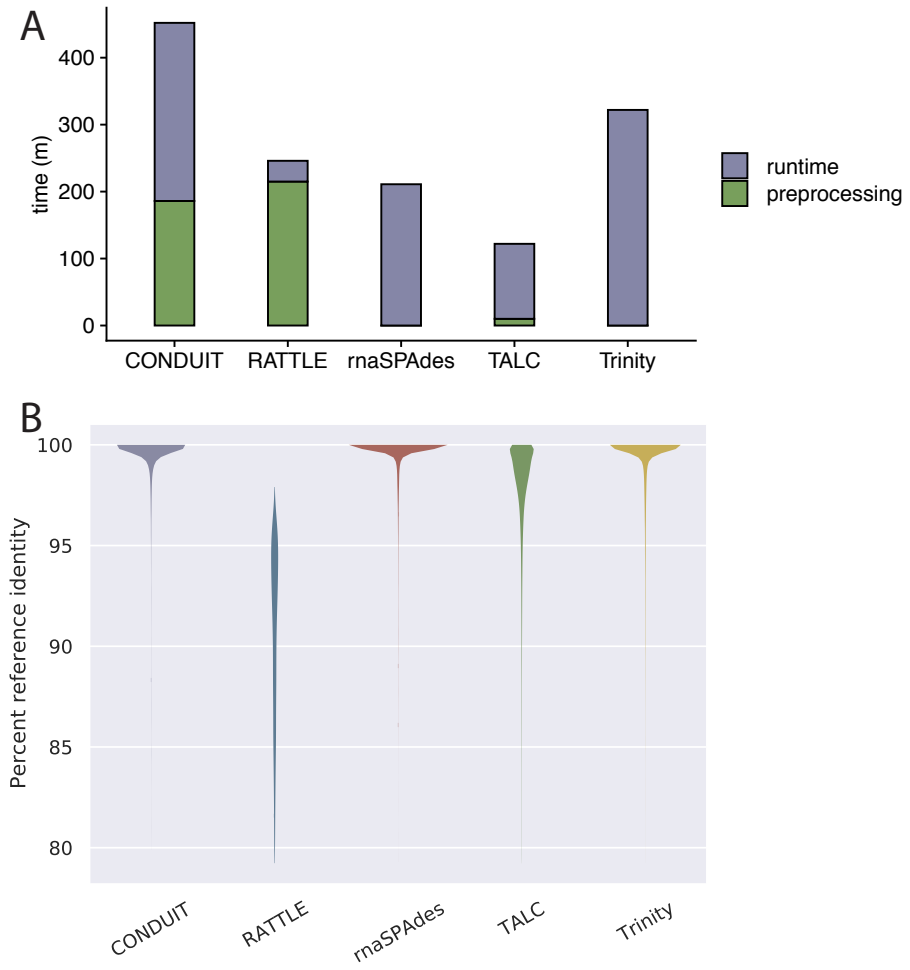


Figure 3.3: Timing and identity benchmarking of CONDUIT. **(A)** Real time taken for preprocessing and runtime for various tools running on the *C. elegans* L4 dataset running on 40 threads. **(B)** Percent reference identity metric of extracted isoforms from various tools upon alignment to the ce11 reference genome.

for each tool evaluated is summarized in Supplemental Table 2. As can be seen, CONDUIT produces isoforms with a high degree of correspondence to the reference genome (median 99.9%), with a comparable degree of percent reference identity to Trinity and rnaSPAdes (median 100% for both), and a greater percent reference identity than RATTLE based assembly (median 91.6%) and TALC based correction (median 98.7%). Indeed in the *C. elegans* L4 dataset 29.6% of CONDUIT called isoforms exactly matched the reference genome at the nucleotide level according to the "NM" tags reported in the BAM alignments, compared with 40.8% for Trinity, 45.5% for rnaSPAdes, 0% for RATTLE, and 8.1% for TALC.

We next evaluated the precision and recall of intron chains identified from each algorithm, as calculated by GFFcompare (Figure 3.4A) (Pertea and Pertea, 2020). As can be seen, CONDUIT run with default settings is competitive with RATTLE at improving intron chain precision / recall, and outcompetes Trinity, TALC, and rnaSPAdes in intron chain precision (though has lower recall). In the L4 dataset CONDUIT outperforms the *ab initio* method StringTie2 in intron chain precision as well, though this is not the case in the Male dataset. CONDUIT reports the number of individual long-reads that contributed to each representative isoform, allowing users to filter for representative isoforms with a desired level of long read support. When CONDUIT output is filtered to include representative isoforms with 5 or more reads, the corresponding intron chain level precision dramatically improves > 20% in all *C. elegans* datasets, although this causes the recall to drop substantially in several *C. elegans* datasets. When filtered in this manner CONDUIT has higher precision

than all other tools evaluated in every *C. elegans* stage S3.2. When transcript sequences are sufficiently accurate they can be used to predict the protein produced by that transcript by searching for the longest open reading frame in the transcript and translating that ORF *in silico*. To evaluate the utility of CONDUIT in generating transcripts capable of being used to predict the proteome of an organism *de novo* we generated such predictions and evaluated these predictions relative to annotated proteomes. Using the annotated proteome as a truth set we then determined precision and recall of protein predictions for each tool evaluated (Figure 3.4B). As can be seen CONDUIT outperforms existing *de novo* assembly tools in one or both metrics for every *C. elegans* dataset evaluated, though it is outcompeted in both metrics by the *ab initio* method StringTie2 in several datasets. When *C. elegans* isoforms are filtered for isoforms with 5 or more reads supporting them the precision of these predictions is even higher (87.8% in the L4 stage).

Upon aligning CONDUIT *C. elegans* transcriptome assemblies to the ce11 genome using minimap2 and comparing these assemblies to the reference using GFFcompare we find thousands of novel intron chains not previously included in the annotation (Figure 3.5) (Li, 2018; Pertea and Pertea, 2020). Due to the 3′ bias present in ONT sequencing, it is possible that some of these "novel" intron chains are due to 5′ truncations in sequencing. Supporting this possibility, in the L3 stage of *C. elegans* 2,223 of the novel intron chains were classed by GFFcompare as "contained in reference," meaning they contained a set of introns consistent with a transcript in the reference but did not match this transcript completely. However, some of these novel isoforms were the

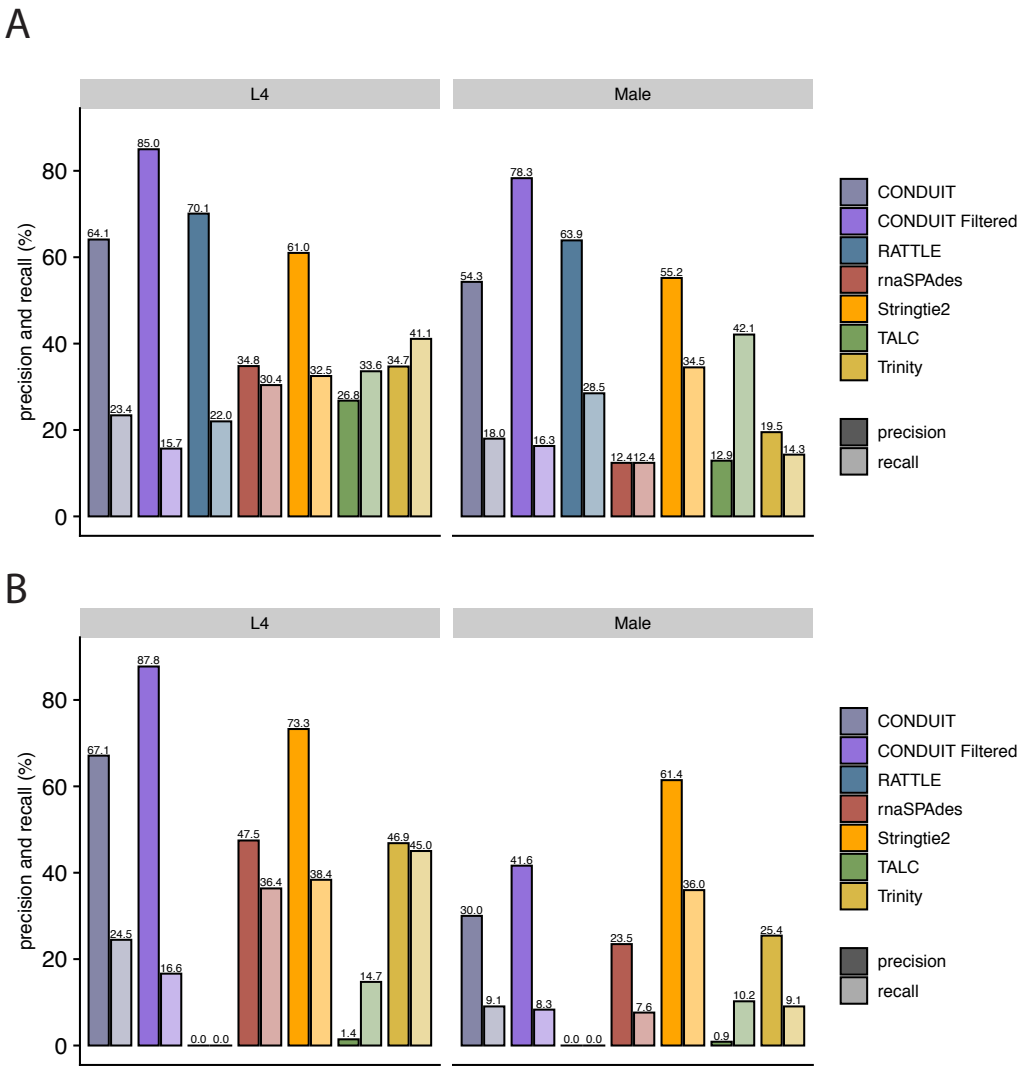


Figure 3.4: Precision and recall of **(A)** intron chains and **(B)** predicted protein products for *C. elegans* L4 and Males for each tool evaluated.

result of identifying novel introns (Figure 3.5B) and novel exons (Figure 3.5C), suggesting CONDUIT is a useful tool for expanding out understanding of the transcriptome through identifying novel exons and introns.

We also evaluated intron chain level (Figure 3.6A) and protein prediction level (Figure 3.6B) precision / recall in data collected from the human B-lymphocyte cell line GM12878 (Workman et al., 2019). Notably we did not use the full dRNA-seq dataset available from this work as the runtimes for a dataset consisting of 30 flowcells worth of dRNA-seq was prohibitively long; we instead opted to use the data produced by a single flowcell from this dataset. Similar to the C. elegans datasets the GM12878 intron chain level precision / recall shows CONDUIT's raw output being less precise than RATTLE and StringTie2 under default settings, but filtering for transcripts with more than 5 reads brings CONDUIT to intron chain precision levels rivaling that of StringTie2 and outcompeting RATTLE, (at the expense of some recall). Protein prediction precision / recall in GM12878 shows that CONDUIT by default outcompetes all *de novo* assembly methods in precision. Recall levels were low for every tool evaluated, likely due to using the entire human transcriptome annotation and proteome as a truth set, when only a subset of the transcriptome and proteome is likely to be expressed in GM12878.

3.5.3 *De novo* transcriptome assembly of *C. nivariensis*

Finally, to show the efficacy of CONDUIT in assembling transcriptomes in the absence of a reference genome, we assembled the transcriptome of *Candida nivariensis*, a pathogenic yeast without a published reference genome, using

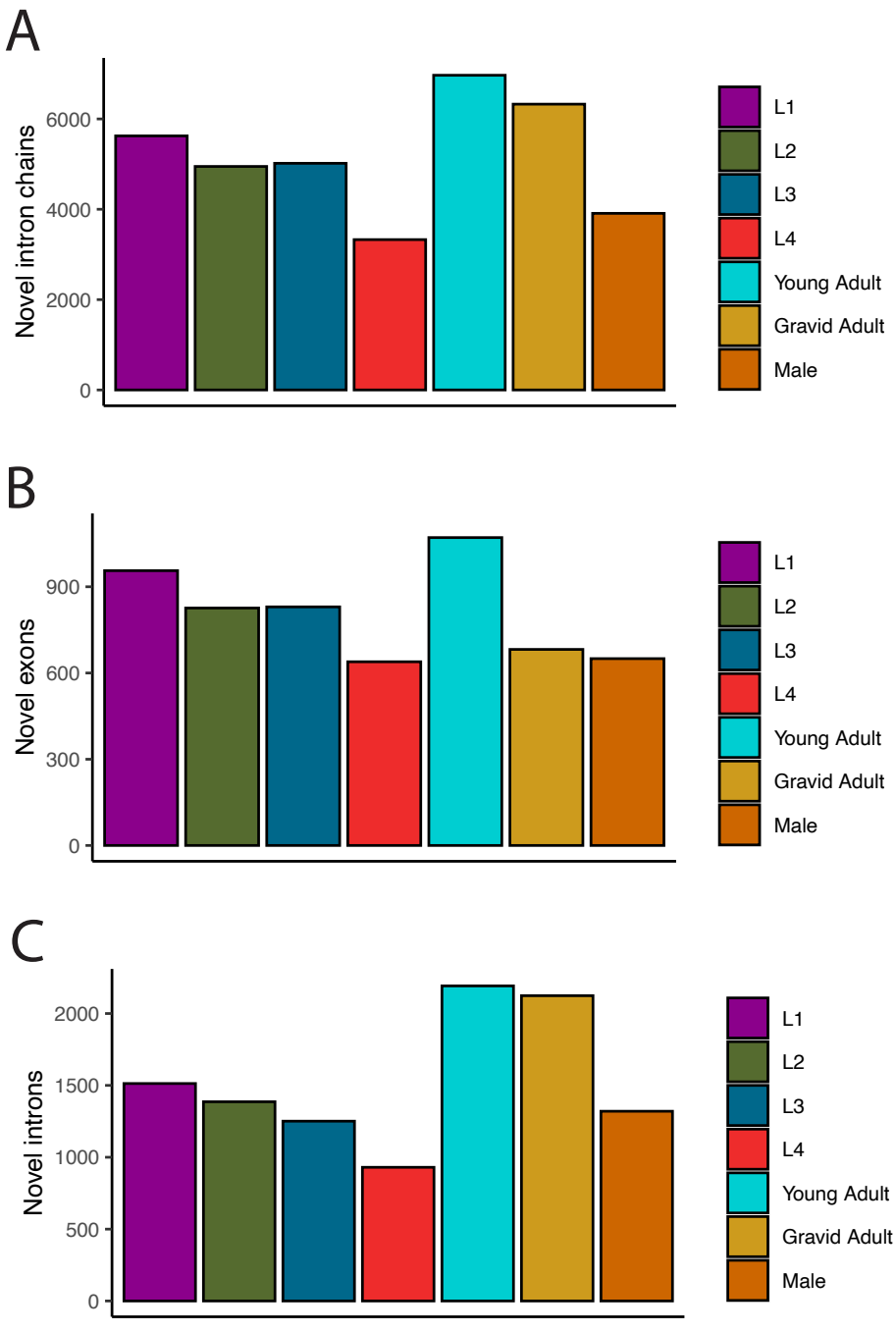


Figure 3.5: Precision and recall of **(A)** intron chains and **(B)** predicted protein products for GM12878 data for each tool evaluated.

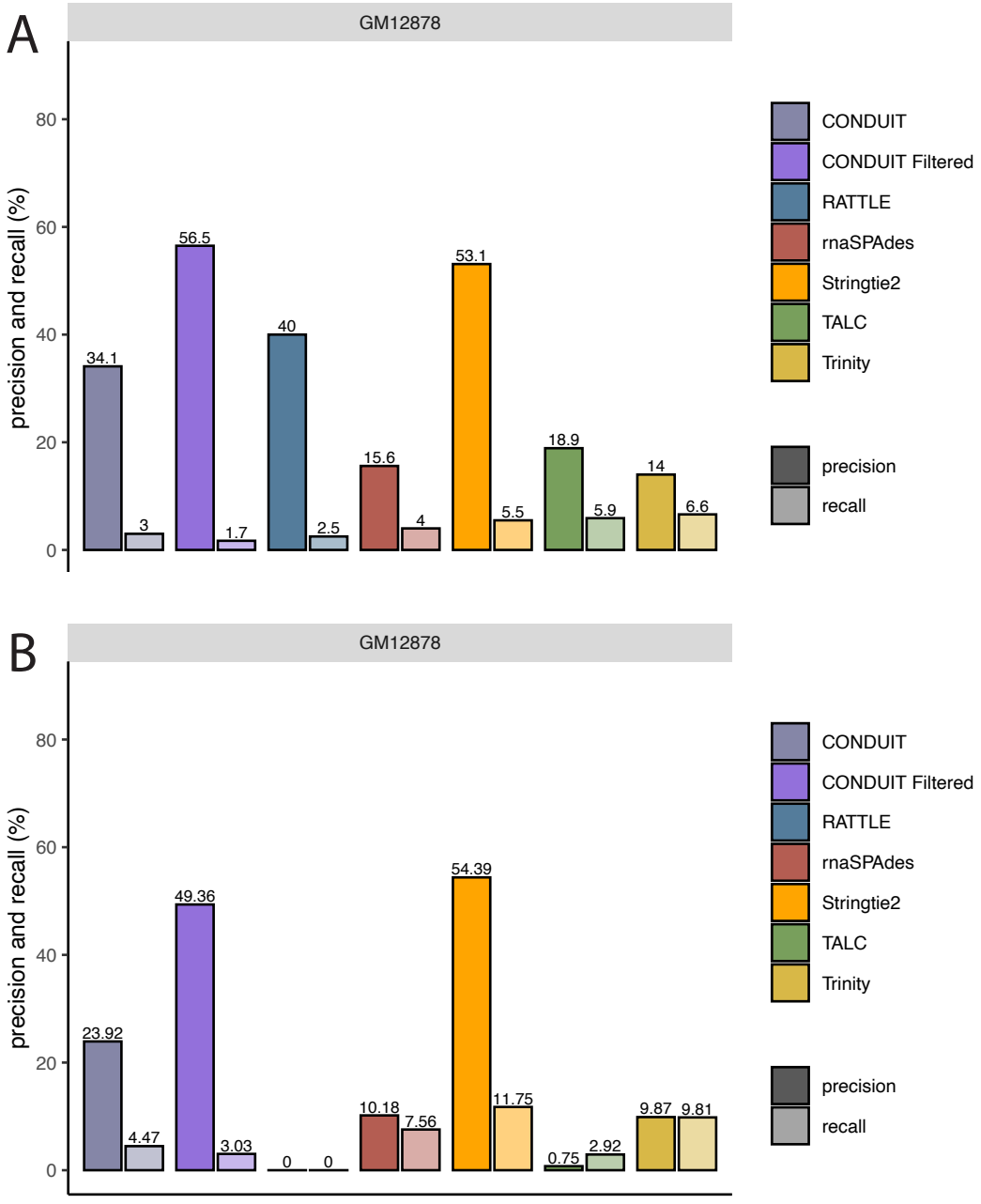


Figure 3.6: Precision and recall of **(A)** intron chains and **(B)** predicted protein products for GM12878 data for each tool evaluated.

ONT based dRNA-seq and corresponding paired-end Illumina RNA-seq. In total CONDUIT generated 17,856 putative isoforms from 5,941 genes with Illumina data supporting every internal base in each isoform (with 100 nt of tolerance at the 5′ and 3′ end of the isoform). This is a higher ratio of putative isoforms to putative genes than one would expect from a yeast closely related to *Candida glabrata*, a species in which there is very little splicing and the ratio of genes to isoforms is close to 1:1 (Linde et al., 2015). Thus many of the isoforms produced in the *Candida nivariensis* data may be erroneous or redundant. Of these putative isoforms 6,941 contained open reading frames longer than 75 amino acids. 233 of these ORF did not match any protein annotated in *Candida glabrata* according to a BLASTP (protein BLAST) search of the predicted proteins against the *Candida glabrata* proteome with an E-value threshold of 1e-10 (Altschul et al., 1997; Schäffer et al., 2001). *Candida glabrata* is the most closely related species to *Candida nivariensis* with a known reference genome, therefore these 233 ORFs likely reflect either errors generated by CONDUIT or newly evolved proteins in *Candida nivariensis* (Alcoba-Flórez et al., 2005).

As an example of the utility of reference-free protein prediction we looked at a *nivariensis* protein match to the *glabrata* gene PDR13, a gene predicted to be involved in pleiotropic drug resistance. Notably this gene appears to be highly conserved between these two species, retaining 90% amino acid sequence identity and containing no insertions or deletions relative to the *glabrata* gene, suggesting an important functional role for this protein (Supplemental Figure S3.4). To determine how a CONDUIT assembly of nivariensis

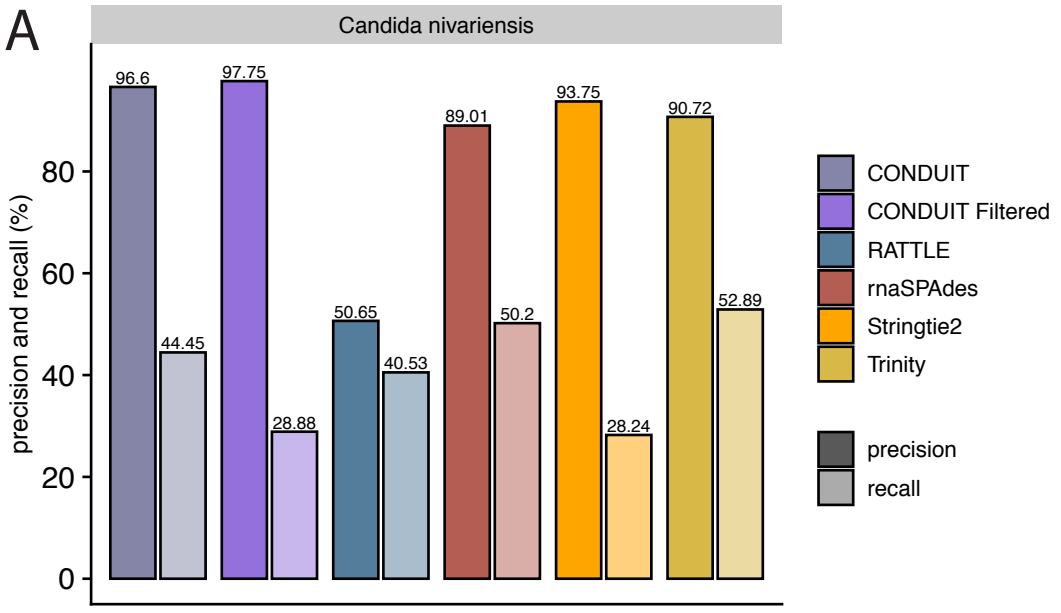


Figure 3.7: (A) Precision and recall of predicted protein products for *Candida nivariensis* data as determined by BLASTP matches against the *Candida glabrata* proteome for each tool evaluated.

compared with other methods for assembling transcriptomes, we compared the CONDUIT assembly with assemblies generated from *de novo* methods rnaSPAdes, RATTLE, and Trinity, and the *ab initio* method StringTie2. To generate the StringTie2 assembly we made use of a draft reference genome for *Candida nivariensis* produced by ONT and Illumina sequencing, assembled with canu under default settings (with the genomeSize parameter set to 12m) (Koren et al., 2017). ONT and Illumina data were aligned to this reference genome using minimap2 and HISAT2 respectively, and these alignments were used to generate a StringTie2 assembly (see Methods) (Li, 2018; Kim et al., 2015; Kovaka et al., 2019). To evaluate these assembled transcriptomes we again made use of precision recall of predicted protein products, this time using the *Candida glabrata* proteome as a truth set and identifying matches by performing a BLASTP search using only the *glabrata* proteome as the target database and the predicted proteins as a query. Matches with an E-value of less than 1e-10 were used to calculate precision and recall. Since multiple predicted proteins can match the same *glabrata* protein we used two definitions of true positives. When calculating precision true positives were defined as the number of predicted proteins that matched a *glabrata* protein, and when calculating recall true positives were defined as the number of *glabrata* proteins that were matched. Predictions were called false positives if they matched no *glabrata* protein, and *glabrata* proteins were called false negatives if no predicted protein matched with them. Using this approach, CONDUIT obtained the highest precision and recall of the tools evaluated, 91% and 44.4% respectively (Figure 3.7). Precision was improved to over 96% when representative isoforms were filtered to include only those isoforms with 5 or more reads

supporting them. CONDUIT also outperformed StringTie2 in these metrics for Candida nivariensis, seemingly in part due to the fact that the reference genome used to generate protein predictions for StringTie2 alignments still contains a number of indels within gene regions, based on Illumina RNA-seq data aligned to the reference genome (Supplemental Figure S3.5).

3.6 Discussion

Here we have demonstrated the utility of our novel approach for *de novo* transcriptome assembly. This approach outcompetes existing *de novo* assembly tools in the task of generating transcripts polished to a level such that protein prediction is possible. Particularly in the *C. elegans* larval and young adult datasets, CONDUIT dramatically outperforms existing tools in the task of protein prediction, with a precision of above 54.5% and recall of the entire annotated *C. elegans* proteome of above 24%.

The quality of transcriptomes assembled by CONDUIT depends on several factors, including the complexity of the transcriptome being reconstructed, the length of the corresponding short read RNA-seq data, and the depth of sequencing of both the long read RNA-seq data and the corresponding short read RNA-seq data. In *C. elegans* L4 stage data, where CONDUIT had it's best performance, the corresponding short read RNA-seq was stranded paired-end 2x150 bp Illumina RNA-seq. In contrast, the *C. elegans* male short read sequencing data was stranded 1x51 bp Illumina RNA-seq. This is notable as the quality of the resulting transcriptomes for these two samples were dramatically different, with a precision of protein prediction in the male data

of roughly half that of the L4 stage. Part of this could be explained if the male proteome of *C elegans* is less well annotated than the proteome of the hermaphroditic larval stages. However the corresponding median percent reference identities in L4 vs male stages (99.9% vs 99.3%) indicates that Illumina polishing was less effective in males, suggesting longer, paired-end Illumina RNA-seq outperforms shorter single-read sequencing at polishing. The intron chain and protein level precision for GM12878 samples, meanwhile, was substantially lower than that of the *C. elegans* L4 sample. This is partially a function of Illumina read lengths (the GM12878 data used 2x101 bp reads), but likely also related to the complexity of the transcriptome being assembled.

CONDUIT's ability to detect novel intron chains, exons and introns indicate that CONDUIT is capable of improving the annotation of even the well annotated *C. elegans* transcriptome. The highly polished nature of the transcripts reported by CONDUIT and the requirement that every internal node and edge of the graph representation of every reported isoform is supported by Illumina RNA-seq allows us to report these novel exons and novel introns with a high degree of confidence.

Though CONDUIT outperformed several existing tools in terms of correspondence to the reference genome, and precision and recall of intron chains and predicted protein products, it was slower than existing long-read correction and *de novo* transcriptome assembly tools. This runtime can likely be improved with some modifications to the partial order alignment step. Several SIMD (Single Instruction Multiple Data) implementations of partial order alignment

have been written, and using such an implementation may speed partial order alignment of clusters significantly (Vaser et al., 2017; Gao et al., 2020). It is worth noting however, that the runtime of CONDUIT scales poorly as the number of long read scaffolds increases, particularly when the preprocessing RATTLE gene level clustering is considered. This poor scaling resulted in very long runtimes (> 1 week on 20 cores) when attempting to run the full dRNA-seq dataset available for GM12878.

CONDUIT outperformed all other tools evaluated at reconstructing the proteome of the pathogenic yeast *Candida nivariensis*, when using BLASTP matches to evaluate precision recall and using the *Candida glabrata* proteome as a truth set. This demonstrates the utility of using CONDUIT when reference genomes are not available, or only a low quality draft genome. It is worth noting however that Illumina RNA-seq coverage for this transcriptome was relatively low, and the Illumina data was not stranded, which likely impacted the performance of de Bruijn graph based methods Trinity and rnaSPAdes.

3.7 Methods

3.7.1 Software availability

CONDUIT is written in Nim and available at `https://github.com/NatPRoach/conduit` under the GPLv2 license. Scripts used in the benchmarking, analysis, and figure generation are available at `https://github.com/NatPRoach/conduit_publication_scripts`.

3.7.2 Data availability

All benchmarking performed in this publication was based on publicly available data generated in previous studies. Supplemental Table 3 contains a summary of metadata regarding each dataset including the publication of origin, where it can be accessed, and accession number (where relevant).

3.7.3 Gene level clustering

For the purposes of clustering long read RNA-seq data into gene level clusters we use the RATTLE clustering algorithm, which based on the RATTLE publication outperforms minimizer-based read clustering (de la Rubia et al., 2020). Briefly, RATTLE computes the clusters by a greedy two step clustering approach. Reads are first compared to one another by comparing the number of common k-mers between the two reads, and if they are similar enough by that metric are then compared using an approach based on determining the Longest Increasing Subsequence of co-linear k-mers between the two reads. Crucially, RATTLE splits long read RNA-seq reads into robust gene level clusters, which is necessary for our hybrid approach.

3.7.4 CONDUIT Partial Order Graph Buildup

Partial order alignment is a means of generating multiple sequence alignments in a data structure known as a partial order graph (Lee et al., 2002). This partial order graph represents the relationship between input sequences as a directed acyclic computational graph where each vertex / node in the graph represents a nucleotide and the edges in the graph represents paths taken through the

graph by one or more sequence reads. In the case of our partial order graph representation the weights for each edge are proportional to the number of reads whose path through the graph contains that edge. The likelihoods for each edge are calculated as the weight of that edge divided by the total sum of weights of edges with the same tail node as that edge, where a tail node is the node from which a directed edge originates.

In the zero-th iteration in which partial order graphs are built, pruned, and first pass representative isoforms are extracted from each gene cluster, the partial order graph construction is performed in a manner that reduces the computational complexity of the problem. Clusters containing more than 200 reads are split into smaller subclusters containing a max of 200 reads, partial order graphs are then constructed, pruned, and representative isoforms are extracted for each subcluster. The resulting subcluster representative isoforms are then progressively merged and decomposed, starting with the subclusters with the fewest number of representative isoforms. After each merge a new partial order graph is calculated for the merged subcluster. The graph is then pruned, and representative isoforms are extracted from the merged subcluster partial order graph (see Methods below). Merging continues until there is only one subcluster left, at which point the extracted representative isoforms are stored for use in later rounds of iterative short read polishing.

CONDUIT uses the software poaV2 (https://github.com/tanghaibao/bio-pipeline/tree/master/poaV2) to perform its partial order graph construction via partial order alignment. poaV2 is provided under the GPLv2 license and was modified slightly for use in CONDUIT. Note that the original

poa software was written by Christopher Lee, and was also distributed under the GPLv2 license (Lee et al., 2002).

3.7.5 CONDUIT Partial Order Graph Pruning

Once partial order graphs for each gene cluster have been generated, CONDUIT prunes these graphs to remove edges and nodes resulting from putative errors in the sequences. Pruning is performed in each iteration of the algorithm, including the first pass isoform extraction step in which only long reads are considered. Pruning is a necessary step to reduce the number of branches in the graph, generate consensus sequences for common regions between isoforms, and to reduce the number of unique isoforms extracted in later steps. This pruning operates under the following two principles and assumptions:

- Reads must have a region of continuous difference of at least Ψnucleotides to be considered distinct isoforms, where $\Psi$is a parameter set by the user (default: 35).
- A single scaffolding read is sufficient to declare the presence of a new isoform. Note that while single reads are used internally to declare new isoforms, final reported isoforms can be filtered after the program is done running based on their degree of long read support.

Pruning is performed in the following manner:

Starting from the start node of each read, a greedy walk through the graph is generated, walking down the edges that have the highest level of support

until either i) a node visited in a previous greedy walk is reached, at which point the walk continues for another Ψnucleotides and then stops, or ii) a node with an outdegree of zero is reached.

Each read is then compared against each greedy walk, and continuous regions of the read that differ relative to the greedy walk by less than Ψnodes are modified to reflect the greedy walk consensus. Continuous differences larger than Ψnodes are assumed to reflect a different isoform and are not modified.

Once each read has been modified to reflect the greedy walks the graph is reconstructed using the corrected reads and each greedy walk is considered one more time. This time, every edge originating from every node present in the greedy walk is considered. The edges that are not a part of the greedy walk are added to a priority queue, ordered by i) the weight of the edge in question ii) the log likelihood of the edge in question iii) the node id of the tail node of the edge iv) the node id of the head node of the edge. This priority queue therefore stores the paths remaining in the graph after an initial greedy pruning, ordered by the degree of support for that edge.

Edges are popped off the priority queue in order. If this edge or the head node of the edge no longer exists in the graph due to a previous trimming step the edge is ignored, otherwise a greedy walk is generated starting with that edge and stopping in the same scenarios described in step 1. As the greedy walk is generated, branches not taken in the greedy walk are added to the priority queue. Each read is then corrected relative to this greedy walk following the same procedure used for earlier greedy walk correction. This process is repeated until the priority queue is empty.

This approach results in a pruned partial order graph that maintains differences between reads that are larger than some parameter Ψ, while collapsing and pruning differences between reads smaller than this Ψparameter. Once the graph is pruned, representative isoforms can be extracted from the graph.

3.7.6 CONDUIT Representative Isoform Extraction

CONDUIT extracts splice isoforms from its trimmed splice graphs using an approach similar to how the program StringTie does (Pertea et al., 2015; Kovaka et al., 2019). Briefly, the position in the splice graph with the highest degree of read support from the scaffolding nanopore reads is selected as a starting point for isoform extraction. Once this point is selected, the isoform is extended greedily until it can no longer be extended. Each greedy extension is performed by adding to the isoform path the edge most supported by corrected reads consistent with the path walked up to that point. In this approach, a single difference between the isoform path and the path of a read is deemed sufficient to remove that read from contributing to the weights considered in the greedy extensions. Once the isoform path is completely extracted, all reads consistent with the full isoform path are removed from the graph such that they no longer contribute to the weights of the graph. This process is repeated until all reads are accounted.

3.7.7 CONDUIT Illumina Polishing

Once first pass nanopore-only consensuses are extracted from each gene cluster, higher quality, shorter reads are incorporated into the approach to

polish the resulting isoforms. To perform this polishing, short reads are aligned using Bowtie2 (Langmead and Salzberg, 2012) to the representative isoforms extracted from the first pass consensus gene models, optionally constricting alignment based on the strandedness of the long read scaffolds and corresponding short read RNA-seq. The partial order graph representation of the relationships between isoforms in a gene cluster is then recalculated using poaV2 (https://github.com/tanghaibao/bio-pipeline/tree/master/poaV2) (Lee et al., 2002). The alignments of the shorter Illumina reads to the scaffolding representative isoforms are combined with the known path through the graph each representative isoforms takes, effectively aligning Illumina reads to the graph. Insertions, deletions, and mismatches in the Illumina reads relative to the scaffolding representative isoform to which they are aligned is used to update the partial order graph by creation of new nodes, new edges, and updating of the weights of edges supported by these Illumina reads. Note that the weight contributed by each Illumina read is a parameter set by the user (default: 10). Once the graph is completely updated with the weights, edges, and nodes supported by the Illumina reads, partial order graph trimming and read correction is performed as above using the updated graph structure. Representative isoforms are then again extracted from the trimmed splice graph and stored. This process of Illumina polishing can be repeated several times, which typically results in improved consensus accuracy with each iteration.

There are two principal reasons for recalculating the partial order graph in each iteration: 1) isoforms that were once different enough to be considered

separate isoforms are occasionally merged within the graph if they converge to the same sequence after polishing; and 2) since we align with Bowtie2's default reporting mode in the graph-based polishing iterations, each short read is aligned to at most one representative isoform. By considering all representative isoforms in their partial order graph representation, the common regions between isoforms will receive support from each short read that aligns in those regions, thereby correcting all the representative isoforms in the graph structure while only aligning to one of them in the linear alignments.

3.7.8 Final polishing and stringent filtering

After the rounds of graph based polishing are performed, CONDUIT includes an optional but recommended round of linear polishing in which each isoform in a cluster is polished separately. This round of linear polishing is included to deal with rare errors that remain after graph based polishing due to interactions between the graph based structure and left alignment of indels in Bowtie2 alignments. By polishing in a linear manner these errors can be polished away.

As a part of this final polishing step a stringent filter is applied to ensure that the reported isoforms are supported by Illumina reads throughout their length. This filter requires that every base and every base to base transition in a reported representative isoform is supported by a short read, with some tolerance at the ends of an isoform to deal with lower rates of Illumina alignments at these end positions. Thus every isoform reported in the default stringent mode of CONDUIT has Illumina reads supporting all internal positions of the

isoform.

3.7.9 Isoform Alignment

To improve stranded correspondence to the reference genome annotation for De Bruijn graph based approaches, extracted isoforms from these approaches were reverse complemented before alignment if the longest open reading frame found in the isoform was found in the reverse complement of the isoform. Isoforms were aligned to the relevant genome (hg38 or ce11) using minimap2 version 2.17-r941 with the following settings "-ax splice" for default alignments, "-ax splice -C 2" for alignments used to identify non-canonical splice sites supported by our approach (Li, 2018). Alignments were converted from SAM to BAM format using samtools version 1.9 (using htslib 1.9) (Li et al., 2009). Alignments were then converted from BAM format to BED format using the BEDTools bamtobed function (using BEDtools version 2.29.2) (Quinlan and Hall, 2010),(Quinlan, 2014). BED files were then converted to GTF format using a custom nim function (bed2gtf) included in the CONDUIT Github repo (https://github.com/NatPRoach/conduit) in the conduitUtils.nim file. This conduitUtils.nim file can be compiled as a command line utility, containing the bed2gtf utility as well as several other utility functions used in our analysis. These GTF files were then compared against a reference transcriptome of the relevant genome using GFFcompare to evaluate metrics of precision and recall at various scales of the transcriptome including base, exon, intron, and intron chain levels (Pertea and Pertea, 2020). The GTF transcriptome references were obtained from the UCSC table browser (Karolchik et al., 2004) (for the human reference the settings were: group: "genes and gene predictions", track:

"GENCODEv32", table: "known gene") (for the C. elegans ce11 reference the settings were: group: "genes and gene predictions" track: "WS245 genes", table: "ws245genes"). The relevant results of GFFcompare for each tool and dataset evaluated can be found in Supplementary Table 2.

3.7.10 Protein Prediction

Protein products were predicted from each extracted representative isoform by searching each isoform for the longest putative ORF, and translating that ORF in silico. Predicted protein products were then filtered to consider only products 75 or more amino acids in length. These predicted protein products were then compared to a database of annotated protein products for the purposes of evaluating precision and recall (the human proteome used was downloaded from GENCODE [the file gencode.v33.pc_translations.fasta.gz] and can be found here: ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_33/ [Frankish et al. (2019)]. The C. elegans proteome used was downloaded from Uniprot and can be found here: https://www.uniprot.org/uniprot/?query=organism:elegans&fil=proteome%3AUP000001940+AND+organism%3A%22Caenorhabditis+elegans+%5B6239%5D%22# [UniProt Consortium (2019)]). The functions used to translate putative transcripts, and used to compare these translations to a reference are available in the CONDUIT GitHub repo (https://github.com/NatPRoach/conduit) in the same conduitUtils.nim file described above.

3.8 Competing interest statement

N.P.R., and W.T., and M.C.S. were reimbursed for conference fees, travel, and accommodation to speak at events organized by Oxford Nanopore Technologies (ONT). W.T. has two patents licensed to ONT (8,748,091 and 8,394,584).

3.9 Acknowledgements

This paper is dedicated to the memory of James Taylor PhD, Ralph S. O'Connor distinguished professor of Biology and Computer Science at Johns Hopkins University, our colleague, mentor, and friend.

These efforts have been funded in part by NIAID grant R01 AI134384 and was supported in part by National Science Foundation award (DBI-1350041).

3.10 References

Albritton SE, Kranz AL, Rao P, Kramer M, Dieterich C, and Ercan S. 2014. Sex-biased gene expression and evolution of the x chromosome in nematodes. *Genetics* **197**: 865–883.

Alcoba-Flórez J, Méndez-Alvarez S, Cano J, Guarro J, Pérez-Roth E, and del Pilar Arévalo M. 2005. Phenotypic and molecular characterization of candida nivariensis sp. nov., a possible new opportunistic fungus. *J. Clin. Microbiol.* **43**: 4107–4111.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, and Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**: 3389–3402.

Broseus L, Thomas A, Oldfield AJ, Severac D, Dubois E, and Ritchie W. 2020. TALC: Transcription aware long read correction.

De Coster W, D'Hert S, Schultz DT, Cruts M, and Van Broeckhoven C. 2018. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* **34**: 2666–2669.

Frankish A, Diekhans M, Ferreira AM, Johnson R, Jungreis I, Loveland J, Mudge JM, Sisu C, Wright J, Armstrong J, et al.. 2019. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**: D766–D773.

Gao Y, Liu Y, Ma Y, Liu B, Wang Y, and Xing Y. 2020. abPOA: an SIMD-based C library for fast partial order alignment using adaptive band. *bioRxiv* .

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, et al.. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**: 644–652.

Jan CH, Friedman RC, Ruby JG, and Bartel DP. 2011. Formation, regulation and evolution of caenorhabditis elegans 3'UTRs. *Nature* **469**: 97–101.

Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, and Kent WJ. 2004. The UCSC table browser data retrieval tool. *Nucleic Acids Res.* **32**: D493–6.

Kim D, Langmead B, and Salzberg SL. 2015. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**: 357–360.

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, and Phillippy AM. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27**: 722–736.

Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, and Pertea M. 2019. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* **20**: 278.

Langmead B and Salzberg SL. 2012. Fast gapped-read alignment with bowtie 2. *Nat. Methods* **9**: 357–359.

Laver T, Harrison J, O'Neill PA, Moore K, Farbos A, Paszkiewicz K, and Studholme DJ. 2015. Assessing the performance of the oxford nanopore technologies MinION. *Biomol Detect Quantif* **3**: 1–8.

Lee C, Grasso C, and Sharlow MF. 2002. Multiple sequence alignment using partial order graphs. *Bioinformatics* **18**: 452–464.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**: 3094–3100.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, and 1000 Genome Project Data Processing Subgroup. 2009. The sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

Lima SA, Chipman LB, Nicholson AL, Chen YH, Yee BA, Yeo GW, Coller J, and Pasquinelli AE. 2017. Short poly(a) tails are a conserved feature of highly expressed genes. *Nat. Struct. Mol. Biol.* **24**: 1057–1063.

Linde J, Duggan S, Weber M, Horn F, Sieber P, Hellwig D, Riege K, Marz M, Martin R, Guthke R, et al.. 2015. Defining the transcriptomic landscape of candida glabrata by RNA-Seq. *Nucleic Acids Res.* **43**: 1392–1406.

Mangone M, Manoharan AP, Thierry-Mieg D, Thierry-Mieg J, Han T, Mackowiak SD, Mis E, Zegar C, Gutwein MR, Khivansara V, et al.. 2010. The landscape of c. elegans 3'UTRs. *Science* **329**: 432–435.

Niu LG, Liu P, Wang ZW, and Chen B. 2020. Slo2 potassium channel function depends on RNA editing-regulated expression of a SCYL1 protein. *Elife* **9**.

Pertea G and Pertea M. 2020. GFF utilities: GffRead and GffCompare. *F1000Res.* **9**: 304.

Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, and Salzberg SL. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**: 290–295.

Prjibelski AD, Puglia GD, Antipov D, Bushmanova E, Giordano D, Mikheenko A, Vitale D, and Lapidus A. 2020. Extending rnaSPAdes functionality for hybrid transcriptome assembly.

Quinlan AR. 2014. BEDTools: the swiss-army tool for genome feature analysis. *Curr. Protoc. Bioinformatics* **47**: 11–12.

Quinlan AR and Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842.

Rang FJ, Kloosterman WP, and de Ridder J. 2018. From squiggle to base-pair: computational approaches for improving nanopore sequencing read accuracy. *Genome Biol.* **19**: 90.

Roach NP, Sadowski N, Alessi AF, Timp W, Taylor J, and Kim JK. 2020. The full-length transcriptome of c. elegans using direct RNA sequencing. *Genome Res.* **30**: 299–312.

de la Rubia I, Indi JA, Carbonell S, Lagarde J, Mar Albà M, and Eyras E. 2020. Reference-free reconstruction and quantification of transcriptomes from long-read sequencing.

Saito TL, Hashimoto SI, Gu SG, Morton JJ, Stadler M, Blumenthal T, Fire A, and Morishita S. 2013. The transcription start site landscape of c. elegans. *Genome Res.* **23**: 1348–1361.

Schäffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, Wolf YI, Koonin EV, and Altschul SF. 2001. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.* **29**: 2994–3005.

Stark R, Grzelak M, and Hadfield J. 2019. RNA sequencing: the teenage years. *Nat. Rev. Genet.* **20**: 631–656.

Tang AD, Soulette CM, van Baren MJ, Hart K, and others. 2018. Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *bioRxiv* .

Tilgner H, Grubert F, Sharon D, and Snyder MP. 2014. Defining a personal, allele-specific, and single-molecule long-read transcriptome. *Proc. Natl. Acad. Sci. U. S. A.* **111**: 9869–9874.

UniProt Consortium. 2019. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**: D506–D515.

Vaser R, Sović I, Nagarajan N, and Šikić M. 2017. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**: 737–746.

Volden R, Palmer T, Byrne A, Cole C, Schmitz RJ, Green RE, and Vollmers C. 2018. Improving nanopore read accuracy with the R2C2 method enables the sequencing of highly multiplexed full-length single-cell cDNA.

Weirather JL, de Cesare M, Wang Y, Piazza P, Sebastiano V, Wang XJ, Buck D, and Au KF. 2017. Comprehensive comparison of pacific biosciences

and oxford nanopore technologies and their applications to transcriptome analysis. *F1000Res.* **6**: 100.

Wenger AM, Peluso P, Rowell WJ, Chang PC, Hall RJ, Concepcion GT, Ebler J, Fungtammasan A, Kolesnikov A, Olson ND, et al.. 2019. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**: 1155–1162.

Wick RR, Judd LM, and Holt KE. 2019. Performance of neural network basecalling tools for oxford nanopore sequencing. *Genome Biol.* **20**: 129.

Workman RE, Tang AD, Tang PS, Jain M, Tyson JR, Razaghi R, Zuzarte PC, Gilpatrick T, Payne A, Quick J, et al.. 2019. Nanopore native RNA sequencing of a human poly(a) transcriptome. *Nat. Methods* **16**: 1297–1305.

Wyman D, Balderrama-Gutierrez G, Reese F, Jiang S, Rahmanian S, Zeng W, Williams B, Trout D, England W, Chu S, et al.. 2019. A technology-agnostic long-read analysis pipeline for transcriptome discovery and quantification.

Xu Z, Hu Y, Deng Y, Chen Y, Hua H, Huang S, Nie Q, Pan Q, Ma DK, and Ma L. 2018. WDR-23 and SKN-1/Nrf2 coordinate with the BLI-3 dual oxidase in response to Iodide-Triggered oxidative stress. *G3* **8**: 3515–3527.

Yang W, Petersen C, Pees B, Zimmermann J, Waschina S, Dirksen P, Rosenstiel P, Tholey A, Leippe M, Dierking K, et al.. 2019. The inducible response of the nematode caenorhabditis elegans to members of its natural microbiota across development and adult life. *Front. Microbiol.* **10**: 1793.

Zhou L, He B, Deng J, Pang S, and Tang H. 2019. Histone acetylation promotes long-lasting defense responses and longevity following early life heat stress. *PLoS Genet.* **15**: e1008122.

3.11 Supplemental Material

3.11.1 Supplemental Figures

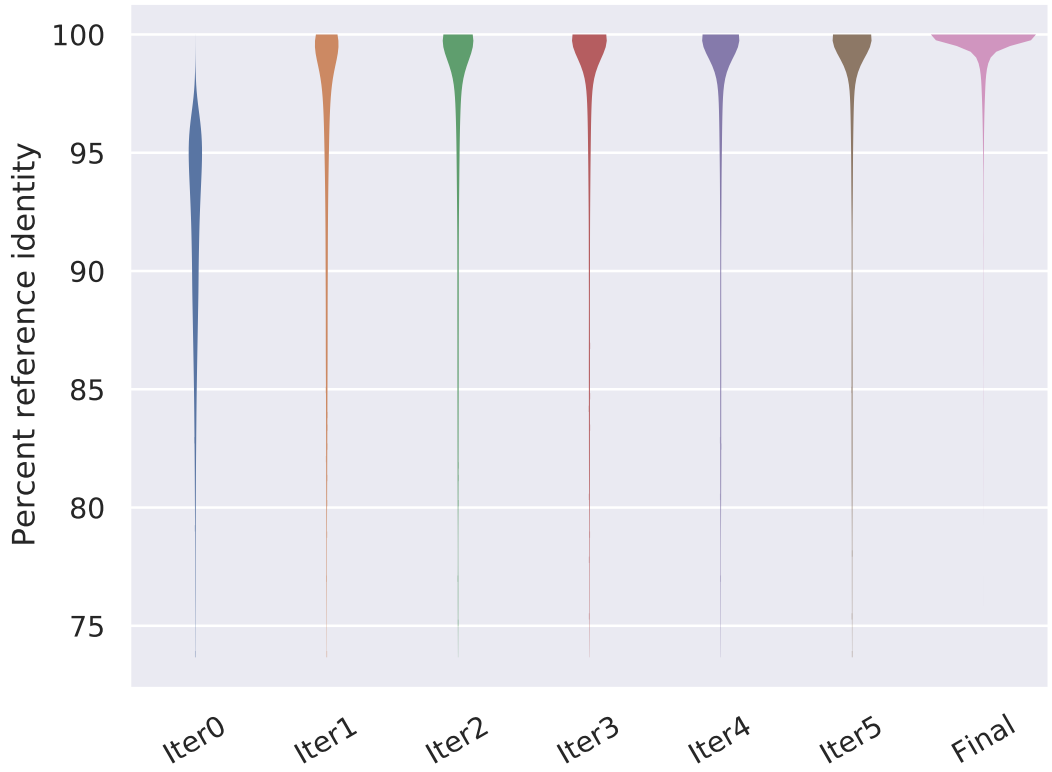


Figure S3.1: Repeated rounds of polishing gradually increase percent reference identity for *C. elegans* L4 data. Stringent threshold in final polish iteration substantially increases average percent reference identity

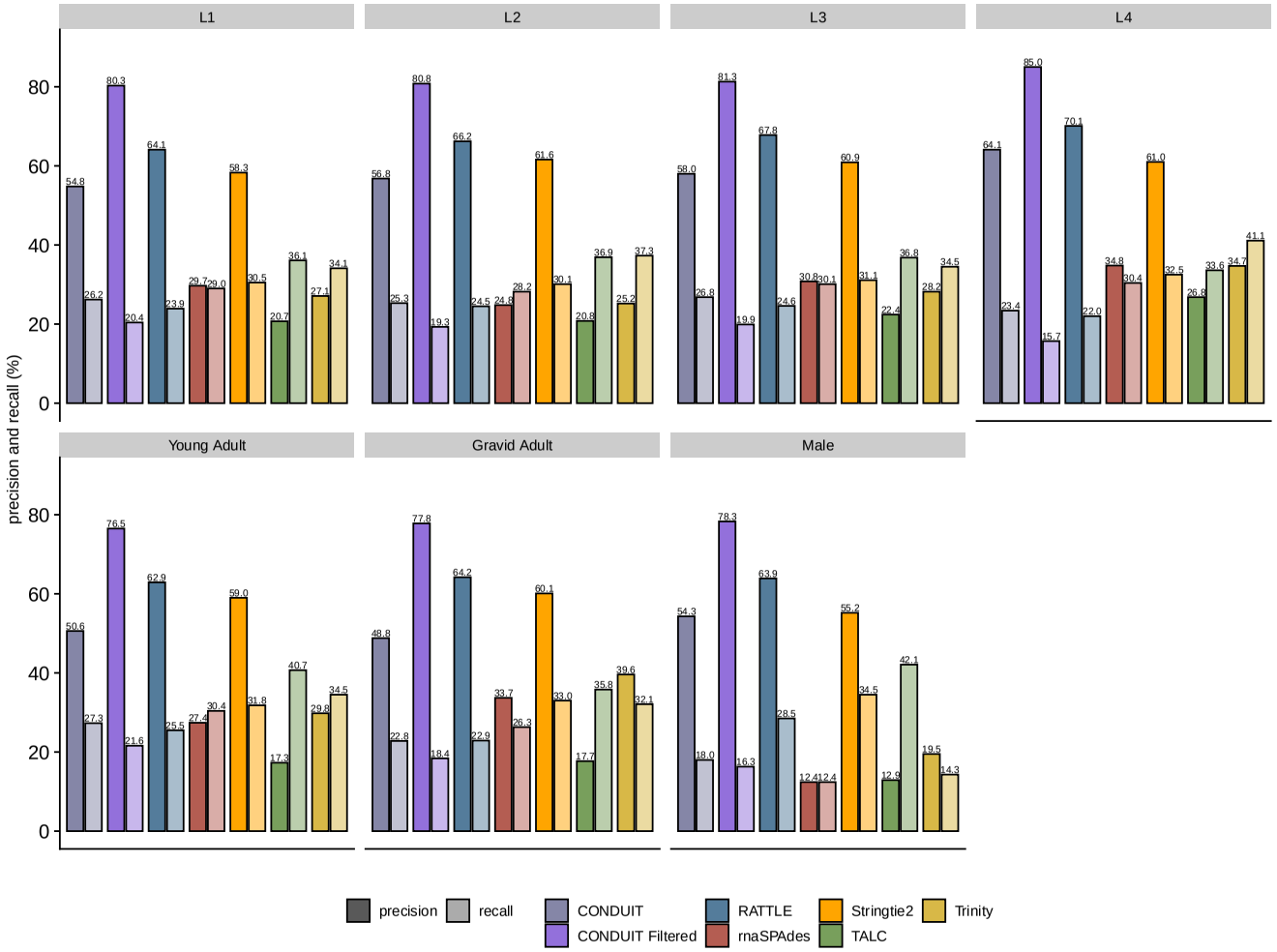


Figure S3.2: Intron chain precision recall in all evaluated *C. elegans* stages for all tools evaluated

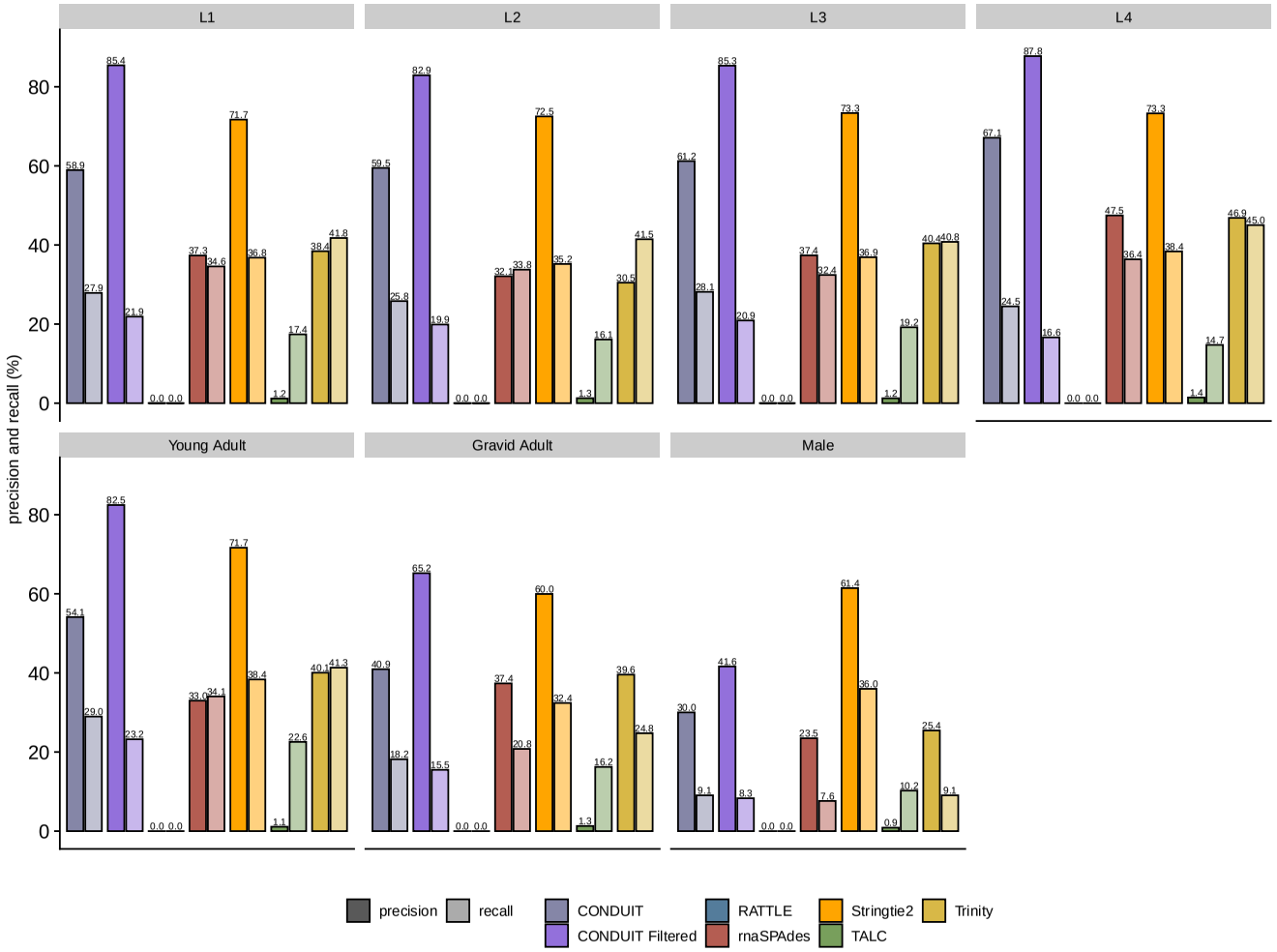


Figure S3.3: Protein prediction precision recall in all evaluated *C. elegans* stages for all tools evaluated

```
Query = cluster_1145_0
Sbjct = CAGL0L10560g PDR13 CGDID:CAL0136038
 Score = 957 bits (2474),  Expect = 0.0, Method: Compositional matrix adjust.
 Identities = 482/536 (90%), Positives = 515/536 (96%), Gaps = 0/536 (0%)

Query  1    MSSPIIGISFGNTSSSIAYINPKNDVDVIANPDGERAIPSVLSYVGEDEYHGGQALQQLV  60
            MSSP+IGISFGNTSSSIAYINPKNDVDVIANPDGERAIPSVLSYVGEDEYHGGQALQQLV
Sbjct  1    MSSPVIGISFGNTSSSIAYINPKNDVDVIANPDGERAIPSVLSYVGEDEYHGGQALQQLV  60

Query  61   RNPKNTIINFRDFIGLPFAQCDVSKCEAGAPVVEIDGKAGFVITRGEKEEKLTVDEVVSR  120
            RNPKNTIINFRDFIGLPF+QCD S+C AGAPVVEIDGKAGFV++RG+ EEKLTVDEVVSR
Sbjct  61   RNPKNTIINFRDFIGLPFSQCDASRCAAGAPVVEIDGKAGFVVSRGDVEEKLTVDEVVSR  120

Query  121  HLNRLKLAAEDYIGSTIPKAVVTVPTNFTEEQKAALKASAAKVGLDVVQFINEPSAALLA  180
            HLNRLKLAAEDYIGS++ +AV+TV ++FT+EQKAALKASAA+VGL++VQFINEPSAALLA
Sbjct  121  HLNRLKLAAEDYIGSSVKQAVLTVSSDFTDEQKAALKASAAQVGLEIVQFINEPSAALLA  180

Query  181  HVEKYPFKEDANVVVADFGGVRSDAAVIAIRNGIFTILATAHDTTLGGDSLDAELIEYFA  240
            HVEKYPFKED NVVVADFGGVRSDAAVIAIRNGIFTILATAHD  LGGD+LDAELIEYFA
Sbjct  181  HVEKYPFKEDVNVVVADFGGVRSDAAVIAIRNGIFTILATAHDQHLGGDNLDAELIEYFA  240

Query  241  KDFEKTNKCNPRKNARALAKLRANAILTKKTLSNATTATISIDSLADGFDYHTSINRMRY  300
            KDFEKTNKCNP KNAR+LAKLRANA++TKKTLSNATTATISIDSLADGFDYHTSINRMRY
Sbjct  241  KDFEKTNKCNPTKNARSLAKLRANAMVTKKTLSNATTATISIDSLADGFDYHTSINRMRY  300

Query  301  ELTANKVFSQFISFIDGVIAKAELDPLDIQAVLLTGGVSFTPKMATNLEFMFPESVEILG  360
            EL ANKVFSQF SFI+ VIAKAELDPLDIQAVLL+GGVSFTPK+ TNLEF+FPESVEILG
Sbjct  301  ELVANKVFSQFTSFIENVIAKAELDPLDIQAVLLSGGVSFTPKLTTNLEFIFPESVEILG  360

Query  361  PQNENATNYPNELNSSGAALQAGLVANYDKDELAEALQPIVLNTPHLPKAIGLVGAHGEF  420
            PQNENA+N PNEL SSGAALQAGLV+NYD +ELAEALQPIV+NTPHLPKAIGL+GAHGEF
Sbjct  361  PQNENASNNPNELASSGAALQAGLVSNYDNEELAEALQPIVVNTPHLPKAIGLIGAHGEF  420

Query  421  HPVLLPETSYPVQKKLTLKNAKGDLLIGVYEGEHHISEKTVEPEAKEADEEDSEEWSDDE  480
            HPVLLPETSYPVQKK+TLKNAKGDLL+GVYEGEHHI EKTVEPEAKE DEEDSEEWSDDE
Sbjct  421  HPVLLPETSYPVQKKITLKNAKGDLLVGVYEGEHHIQEKTVEPEAKEEDEEDSEEWSDDE  480

Query  481  PEVIREKLYTLSTKLMELGVKDVKNGLEIVFNVNKDGVLRVTARDLKSATVVKGEL  536
            PEVIREKLYTLSTKLMELG+KDVKNG EI FNVNKDGVLRV+ARDLKS TVVKGEL
Sbjct  481  PEVIREKLYTLSTKLMELGLKDVKNGAEITFNVNKDGVLRVSARDLKSTTVVKGEL  536
```

Figure S3.4: Example of a BLASTP match between a protein predicted from a *Candida nivariensis de novo* assembly and *PDR13* in the *Candida glabrata* proteome.

Figure S3.5: Evidence of insertions remaining in the *Candida nivariensis* draft reference genome. Blue vertical lines in the reads represent insertions in this IGV screenshot of RNA-seq reads aligning to a locus in the draft genome. These insertions likely affected the ability of StringTie2 to predict proteins.

3.11.2 Supplemental Tables

	Iter0	Iter1	Iter2	Iter3	Iter4	Iter5	Final
Average percent identity:	91	95.4	96.2	96.7	96.9	97.1	99.3
Median percent identity:	92.5	98.3	99.4	99.7	99.8	99.8	99.9

Table S3.1: Repeated rounds of polishing gradually increases percent reference identity for *C. elegans* L4 data. Stringent threshold in final iteration substantially increases average percent reference identity.

Table S3.2: Benchmarking statistics for tools and datasets evaluated, provided as a supplemental .xlsx file to this document

Table S3.3: Accession numbers and citations for datasets used in benchmarking, provided as a supplemental .xlsx file to this document

Chapter 4

Discussion, Conclusions, and Future Directions

4.1 Future directions in *C. elegans* transcriptomics

Though our research has advanced our understanding of the *C. elegans* transcriptome, a great deal remains to be discovered. As advances are made in nanopore sequencing and molecular biology more techniques will become feasible, and the depth of insight that can be gained through RNA-seq will be expanded.

The first and most direct next steps for the research performed in this thesis would be to sequence additional developmental stages not represented in our sequencing. Sequencing of the stress-induced dauer stage, for example, may provide insight into the transcriptional program of this state, the diversity of transcripts expressed in this stage, and may assist in identifying novel isoforms and gene models. Sequencing of embryos at various stages may also prove

useful, as the transcripts present in early embryos are crucial for development and may have novel transcript isoforms that remain to be discovered.

An additional sequencing approach that may be worth pursuing is the sequencing of tissue specific samples. In this approach one can drive expression of a FLAG tagged Poly(A) Binding protein (FLAG::PABP) under various tissue specific promoters. In the same operon as the FLAG::PABP is an SL2 trans-splice site and a GFP, such that the expression profile of the construct can be evaluated through microscopy. A diagram of such a tissue specific FLAG::PABP construct, as well as example GFP images from worms expressing these constructs can be seen in Figure 4.1A.

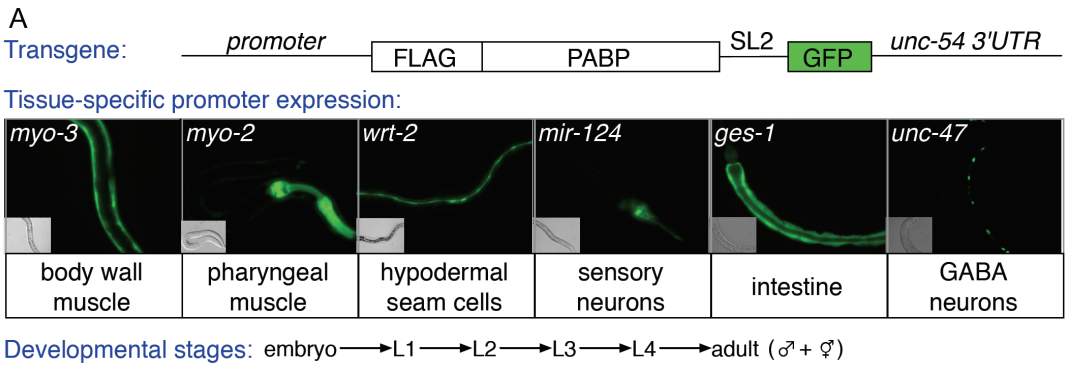


Figure 4.1: Tissue specific PABP expression system proposal for *C. elegans*

Once tissue specific expression has been shown through microscopy the transgene expressing worms are homogenized, an immunoprecipitation for FLAG is performed, (thus enriching the sample for RNAs bound to FLAG::PABP), RNA is extracted, and then sequenced as normal. A similar version of this approach has been used in *C. elegans* to profile 3' UTRs in various tissues, demonstrating the efficacy of the technique. It is possible on a long enough time scale, of course, for the FLAG tagged PABP to unbind RNA from the

target tissue and bind RNA from the rest of the organism, thus steps from homogenization to immunoprecipitation must be performed quickly.

Expanding the number of sequencing datasets collected in *C. elegans* to capture specific tissues and developmental stages not sequenced in our initial experiments should expand our understanding of the transcriptome and bring us closer to the theoretical capture of the complete *C. elegans* transcriptome.

In addition to sequencing of additional developmental contexts of wild-type worms, sequencing of genetic mutants, particularly of known splicing regulators could provide insight into the dysregulation of splicing under various genetic perturbations. By looking globally at differential splicing in these genetic mutants information about the function of the mutated gene could be determined, such as putative binding motifs for the gene's product, identified as motifs enriched in the regions around affected splice sites.

Long read RNA sequencing of wild isolates of *C. elegans* may also provide functional insights into the biology of worms. The *C. elegans* Natural Diversity Resource project (CeNDR), has isolated and DNA sequenced over 700 strains of *C. elegans* from around the world (Cook et al., 2017). By performing long-read RNA sequencing of many of these worms, and characterizing the effect of natural diversity on differential splicing patterns of various genes it is possible that correlations between SNPs and patterns of splicing can be determined. Such data could theoretically be used to train machine learning models to predict RNA splicing patterns from DNA sequence in *C. elegans*, as has previously been done in human cell lines with some success (Jaganathan et al., 2019).

4.2 Comparison with Li et al.

Another research work, published alongside ours, also utilized dRNA-seq to characterize the transcriptome of *C. elegans* in various developmental stages (Li et al., 2020). This work, published in Genome Research by Li and colleagues, sequenced N2 *C. elegans* as embryos, larval stage L1s, and young adults, and based on their analysis claim to support far more novel isoforms in these three developmental stages (sequenced to a depth of approximately 6 million reads total) than supported in our 7 developmental stages (sequenced to a depth of approximately 5.5 million reads total). The discrepancy between the number of reported novel isoforms is likely a factor of the analysis pipeline utilized, and as such, we seek to compare and contrast these methods to determine which is likely a more accurate reflection of the true *C. elegans* transcriptome.

To begin, Li and colleagues state they successfully map 99.7% of their dRNA-seq reads to the ce11 genome, a substantially higher percentage of reads mapped than our study, which mapped only 87.8% of reads. Though the exact parameters of mapping used by Li et al. are not clear, as downloading and mapping their L1 stage data with only the parameters stated in their paper yields only 80% alignment.

Regardless of how mapping was performed, the number of novel isoforms extracted would likely not differ as dramatically between publications if different analysis techniques downstream had not been utilized.

To begin, both Li et al and our research group filter their reads for 'full-length' reads, to deal with the problem of 5' truncations common in long read RNA-seq. Li et al opt to perform this filtering by looking for reads with a splice leader sequence at their 5' end (shortfalls in the analysis of which will be addressed shortly), or by setting an arbitrary threshold of necessary percent coverage of an annotated transcript to be considered 'full-length'. The definition of full-length transcript from Li et al is important as they state they require isoforms be supported by at least 5 independent 'full-length' transcripts to be reported. As such their ability to identify transcripts with SL sequences must be examined thoroughly for any potential problems. They utilize an approach for identifying SL sequences in which they consider the first 22 5'-most nt of each read and perform local sequence alignment against 7 SL sequences. They then define any read with an alignment score above some arbitrary threshold as having an SL sequence at it's 5' end. They defend the validity of this approach and their isoform quantification approach by simulating at a single gene a number of reads corresponding to possible isoforms, then running their analysis pipeline on these simulated reads, and claiming they have low false discovery rates in these simulations. Of course it is not clear whether the gene they chose for these simulations is representative of the transcriptome as a whole in terms of enrichment for SL sequence like motifs and the cutoff they use may therefore be too lenient. In addition, utilizing the first 22 nt of the dRNA-seq reads for SL identification is not appropriate, as it is well documented that the 5' most 10 - 15 bases of each dRNA-seq read are not sequenced (Workman et al., 2019). Thus 10 - 15 of the 22 nt included in their analysis may result in higher rates of false positives.

In addition to this consideration, Li et al opt not to exclude intron retention transcripts from consideration in their isoform definitions. Given that we filter these reads, this explains some of the discrepancy. In addition, we separate out novel UTR isoforms from novel splice isoforms, which Li et al do not do. Given that the vast majority of their novel isoforms are "5′ missing" (which may be a result of falsely labeling reads as containing SL sequences; and which we do not consider), "intron retentions" (which we remove), "UTR extensions", "UTR truncations" (which we do not report as novel isoforms but as a separate count of novel UTRs) "new junction", "5′ extra", or "3′ extra" (which we would not identify due to our restriction of using only existing splice donors and acceptors in the annotation) the discrepancy between number of novel isoforms can be explained.

4.3 Future directions in development of CONDUIT:

Though CONDUIT outperforms existing tools in its current state, there are several avenues for improving its performance and utility. The integration of super-reads, synthetically longer reads generated by stitching together short reads based on their overlapping sequences, for example, may improve the polishing step of CONDUIT (Zimin et al., 2013). These longer and highly accurate super reads would be more capable of resolving regions of representative isoforms with higher error rates and therefore would be able to correct regions that cannot be resolved by short reads alone.

CONDUIT would also benefit from the ability to merge transcriptomes generated in separate sequencing runs, such that common isoforms are identified,

merged, and labeled with the same transcript ID. This is a non-trivial task, as it would require comparison of two transcriptome annotations against each other, which would require a large number of comparisons. Using an approach similar to RATTLE gene level clustering, in which comparisons are broken into two steps, the first extremely fast and the second slower, these comparisons could likely be performed in an acceptable timeframe.

In addition, CONDUIT could be modified to report transcript expression levels for each extracted isoform. Since short reads are already being aligned to representative isoform scaffolds, extracting out transcripts per million (TPM) metrics for each isoform should be relatively computationally inexpensive. Adding this functionality would increase the usefulness of the CONDUIT program as when combined with the merging approach proposed one could in theory perform differential expression analysis between two conditions by quantifying the expression levels in CONDUIT.

Finally, CONDUIT could be modified to output the partial order graphs generated by poaV2 in a manner such that the relationship between isoforms originating from the same cluster could be analyzed and visualized for more insight into each gene cluster. This would allow researchers to examine their genes of interest in it's splice graph format, even in the absence of a reference genome.

4.4 Future directions in the field of long-read transcriptomics

The ability to capture long-read transcripts at full length, without the need to computationally reconstruct transcripts offers new opportunities in the fields of transcriptomics. However the high error rates and various biases associated with these long read technologies require some form of mitigation in order to best utilize this new data type.

The problem of extreme 3′ bias and 5′ truncations, in particular, requires addressing as this problem makes it difficult to impossible to separate out *bona fide* alternative transcriptional start sites from long read RNA-seq data. As proposed in Chapter 2, modifying the long read RNA-seq protocol to enrich for full length transcripts or better identify full-length transcripts in some capacity (either through mild treatment with a 5′ to 3′ exonuclease, pull-down of some 5′ cap associated protein, or tagging the 5′ end with a set of nucleotides of known sequence) would likely greatly facilitate the computational analysis of dRNA-seq reads. By enriching for full-length reads one could be more confident that 5′ truncations observed in the reads reflect alternative transcription start sites, and thereby allow for identification of additional isoforms.

In addition to modifying the molecular biology involved in generating long read sequencing datasets, there is future work to be done in establishing best computational practices for long read RNA-seq. As detailed in Chapter 1, there are a number of transcriptome annotation programs that have been published

since the advent of long read RNA-seq, however all of these programs are quite similar to one another and none of them have established themselves as a clear standard in the field. Given the highly competitive nature of establishing long read RNA-seq pipelines we feel it is unlikely that our ab initio approach, which was tailored to work in *C. elegans*, will be utilized by many in the future. However we feel that the 3′ UTR calling approach we utilized improves on the UTR calling approaches of other published techniques, and believe efforts should be taken to integrate this approach into other long read RNA-seq pipelines more likely to be utilized by other research groups.

Broadly, the application of long-read RNA sequencing to annotation of transcriptomes has thus far mostly been used to expand the annotation of model organisms and human cell lines, with well established and well annotated transcriptomes (Workman et al., 2019; Roach et al., 2020; Volden et al., 2018; Tang et al., 2018; Li et al., 2020; Jenjaroenpun et al., 2018; Garalde et al., 2018). The application of long-read RNA-seq to non-model organisms without well annotated transcriptomes (or even well constructed genomes) is a logical next step now that initial best practices for long read RNA-seq analysis have been established. The application of the techniques established in model organisms to non-model organisms could allow for rapid high-throughput and accurate annotation of transcriptomes at relatively low cost. Performing such analysis at scale in a variety of organisms from the same clade could allow for a long read sequencing based comparative transcriptomic or proteomic studies, which could provide insight into the underlying biology of such a clade. Doing so in a reference free manner, using a tool like CONDUIT would eliminate

the requirement of such a study to have high quality reference genomes for each species sequenced.

There is some evidence that long-read RNA-seq can be quantitative in nature, (though the relatively low throughput as compared to Illumina sequencing makes long-read RNA-seq more prone to drop out of low abundance transcripts) (Wyman et al., 2019). Provided this result is corroborated with additional studies, the application of long-read RNA-seq in an expression quantitative trait loci (eQTL) type study could be highly informative. By correlating genomic SNPs vs the expression level of individual transcript isoforms, and potentially the ratio of isoforms to one another the impact of such SNPs on complicated patterns of splicing prevalent throughout the human genome could be used to inform our understanding of splicing, and our ability to predict splicing patterns from genomic sequence.

Direct RNA-seq is also able to detect RNA modifications at the signal level owing to the impact of these modifications on the electrical signal of the resulting read as it passes through the pore. Several approaches capable of calling these modifications have been developed, though most existing RNA modification detection algorithms focus on N6-methyladenosine (m6A), the most prevalent RNA modification in most organisms (Lorenz et al., 2020; Stoiber et al., 2017). Future work in the field of long-read transcriptomics should make efforts to expand the number of possible RNA modifications detectable from such approaches. This is a non-trivial computational task however, as the number of possible sequences of length k scales with the effective number of bases n, as n^k, limiting the number of modifications one

can consider at any given time. Approaches that compare electrical signals between two samples have also been developed, and are in theory capable of detecting sites of RNA modifications more generally, even if a specific model for that modification has not yet been trained (Leger et al., 2019).

One advantage of long read RNA-seq that has not yet thoroughly been taken advantage of is the ability of reads spanning the full-length of transcript to effectively correlate aspects of transcript structure with one another. For example, there could be instances in which the inclusion or exclusion of a certain exon could correlate with the 3'UTR choice of the transcript. Indeed there are several papers indicating that coupling between transcriptional processing events occur in some genes (Tilgner et al., 2015; Anvar et al., 2018; Tilgner et al., 2018). This was addressed in part in Chapter 2, section X, in which we stated there were few examples of such correlation between splicing and 3'UTRs in our sequencing of the *C. elegans* transcriptome. However, this may be a consequence of relatively low coverage, and increased sequencing depth may provide additional statistical power necessary to detect such correlations. In the event correlations between 3'UTR choice, TSS choice, RNA modifications, and splicing structure can be detected these events could be examined in more detail to determine the underlying molecular mechanism behind such coupling. Examining such coupling in more detail would be non-trivial, and would likely require the establishment of some high throughput screening method to determine where and when such coupling is occuring. In the event a gene was found to have correlated TSS choice and 3'UTR choice, one could attempt a screen in which PP7 hairpins are inserted in the coupled candidate

5'UTR and MS2 hairpins are inserted in the coupled candidate 3'UTR. By then expressing PP7 coat protein and MS2 coat protein tagged with mCherry and GFP respectively, and screening for organisms in which GFP and mCherry signals are non-overlapping one could in theory find candidate genes necessary to drive coupling of 5' and 3'UTR choice in this candidate gene. It is less clear how one would experimentally examine coupling of RNA modifications or splicing patterns within the coding sequence of a gene, as detecting RNA modifications in a high throughput manner without sequencing has not yet been done, and manipulation of the RNA within the CDS would likely disrupt the function of the protein product of such a gene.

4.5 Discussion

The insight into the *C. elegans* transcriptome afforded to us by long and short read RNA-seq has facilitated biological research, increased our understanding of gene regulation, and assisted in gene annotation efforts. However, much remains to be discovered about the nuances of gene regulation, the sets of isoforms expressed by *C. elegans* in various developmental, environmental, and genetic circumstances, and the biology of *C. elegans* more broadly.

In order to obtain these biological insights RNA sequencing data must be analyzed in a robust, reproducible, and logically and statistically sound manner. Ideally this analysis would leverage previously reported data sets and our current understanding of the transcriptome to build on that understanding.

The work outlined in Chapter 2 of this thesis made strides towards both biological understanding of the *C. elegans* transcriptome and establishment

of best computational practices for the handling of long read RNA-seq. We feel our work has advanced the long read RNA-seq field through the creation of a novel means of calling 3'UTRs from dRNA-seq data, and has advanced biological understanding of the transcriptome through profiling of poly(A) tails throughout development, supporting existing transcript isoforms with full-length data, and characterization of novel transcript isoforms.

Meanwhile the development of CONDUIT is a step forward in the field of de novo transcriptome assembly. CONDUIT outperforms existing de novo assembly tools, as shown in Chapter 3. By leveraging the strengths of long reads to generate scaffolds for representative isoforms and the strengths of short reads to polish these isoforms CONDUIT generates highly accurate transcript models completely independently of a reference genome. CONDUIT will therefore be of use to researchers who seek to annotate the transcriptome of an organism without a reference genome or with a poorly constructed reference genome.

4.6 References

Anvar SY, Allard G, Tseng E, Sheynkman GM, de Klerk E, Vermaat M, Yin RH, Johansson HE, Ariyurek Y, den Dunnen JT, et al.. 2018. Full-length mRNA sequencing uncovers a widespread coupling between transcription initiation and mRNA processing. *Genome Biol.* **19**: 46.

Cook DE, Zdraljevic S, Roberts JP, and Andersen EC. 2017. CeNDR, the caenorhabditis elegans natural diversity resource. *Nucleic Acids Res.* **45**: D650–D657.

Garalde DR, Snell EA, Jachimowicz D, Sipos B, Lloyd JH, Bruce M, Pantic N, Admassu T, James P, Warland A, et al.. 2018. Highly parallel direct RNA sequencing on an array of nanopores. *Nat. Methods* **15**: 201–206.

Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, Darbandi SF, Knowles D, Li YI, Kosmicki JA, Arbelaez J, Cui W, Schwartz GB, et al.. 2019. Predicting splicing from primary sequence with deep learning. *Cell* **176**: 535–548.e24.

Jenjaroenpun P, Wongsurawat T, Pereira R, Patumcharoenpol P, Ussery DW, Nielsen J, and Nookaew I. 2018. Complete genomic and transcriptional landscape analysis using third-generation sequencing: a case study of saccharomyces cerevisiae CEN.PK113-7D. *Nucleic Acids Res.* **46**: e38.

Leger A, Amaral PP, Pandolfini L, Capitanchik C, and others. 2019. RNA modifications detection by comparative nanopore direct RNA sequencing. *BioRxiv* .

Li R, Ren X, Ding Q, Bi Y, Xie D, and Zhao Z. 2020. Direct full-length RNA sequencing reveals unexpected transcriptome complexity during caenorhabditis elegans development. *Genome Res.* **30**: 287–298.

Lorenz DA, Sathe S, Einstein JM, and Yeo GW. 2020. Direct RNA sequencing enables m6a detection in endogenous transcript isoforms at base-specific resolution. *RNA* **26**: 19–28.

Roach NP, Sadowski N, Alessi AF, Timp W, Taylor J, and Kim JK. 2020. The full-length transcriptome of c. elegans using direct RNA sequencing. *Genome Res.* **30**: 299–312.

Stoiber M, Quick J, Egan R, Lee JE, Celniker S, Neely RK, Loman N, Pennacchio LA, and Brown J. 2017. De novo identification of DNA modifications enabled by Genome-Guided nanopore signal processing.

Tang AD, Soulette CM, van Baren MJ, Hart K, and others. 2018. Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *bioRxiv* .

Tilgner H, Jahanbani F, Blauwkamp T, Moshrefi A, Jaeger E, Chen F, Harel I, Bustamante CD, Rasmussen M, and Snyder MP. 2015. Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events. *Nat. Biotechnol.* **33**: 736–742.

Tilgner H, Jahanbani F, Gupta I, Collier P, Wei E, Rasmussen M, and Snyder M. 2018. Microfluidic isoform sequencing shows widespread splicing coordination in the human transcriptome. *Genome Res.* **28**: 231–242.

Volden R, Palmer T, Byrne A, Cole C, Schmitz RJ, Green RE, and Vollmers C. 2018. Improving nanopore read accuracy with the R2C2 method enables the sequencing of highly multiplexed full-length single-cell cDNA.

Workman RE, Tang AD, Tang PS, Jain M, Tyson JR, Razaghi R, Zuzarte PC, Gilpatrick T, Payne A, Quick J, et al.. 2019. Nanopore native RNA sequencing of a human poly(a) transcriptome. *Nat. Methods* **16**: 1297–1305.

Wyman D, Balderrama-Gutierrez G, Reese F, Jiang S, Rahmanian S, Zeng W, Williams B, Trout D, England W, Chu S, et al.. 2019. A technology-agnostic long-read analysis pipeline for transcriptome discovery and quantification.

Zimin AV, Marçais G, Puiu D, Roberts M, Salzberg SL, and Yorke JA. 2013. The MaSuRCA genome assembler. *Bioinformatics* **29**: 2669–2677.

Nathan P. Roach
nroach2@jhu.edu https://www.github.com/NatPRoach (716)-270-3529

EDUCATION

Johns Hopkins University
Doctorate of Philosophy (Ph.D.), Biology August 2015 - June 2020 (Expected)
Cell, Molecular, Developmental Biology and Biophysics (CMDB) Department
Laboratories of Dr. James Taylor and Dr. John Kim

The University of North Carolina at Chapel Hill
Bachelors of Science (B.S.) 2015
Biology (with Highest Honors) and Computer Science

RESEARCH EXPERIENCE

Johns Hopkins University
Graduate Student Researcher 2015 - Present
Laboratories of Dr. James Taylor and Dr. John Kim

- Collaborated directly with experimental biologists to determine the impact of various genetic mutants on the transcriptome through analysis of RNA-seq, direct RNA-seq, and small RNA-seq.
- Pioneered a research project utilizing direct RNA-seq from Oxford Nanopore Technologies to profile the transcriptome of *C. elegans*
- Designed and implemented computational approaches to analyze the resulting data, authoring a paper detailing these methods and the results of this analysis.

The University of North Carolina at Chapel Hill
Undergraduate Researcher Spring 2014 - Spring 2015
Laboratory of Dr. Kenneth Jacobson

- Established methodology to characterize the size distribution of nanometer-scale folds of the cell membrane observed after drastic morphological change. The resulting analyses were included in a publication in PLOS Computational Biology.
- Developed and implemented a method to visualize cell migration in three dimensions, through embedding cells in a 3-D collagen matrix

Undergraduate Researcher Fall 2012 - Spring 2014
Laboratory of Dr. Tyson Hedrick

- Developed and implemented computer vision approaches to quantify images of organisms in flight through image segmentation and 3D particle tracking methods

Undergraduate Researcher Fall 2011 - Spring 2012
Laboratory of Dr. Mark Hollins

- Designed and implemented a user interface used for data collection in the lab

Hauptman-Woodward Research Institute
Summer Research Assistant Summer 2013
Laboratory of Dr. Timothy Umland

- Purified and crystalized the bacterial enzyme chorismate synthase to 5Å resolution. Built experience with preparing buffers, sterile technique, several chromatography methods, and gel electrophoresis.

TECHNICAL SKILLS
Programming Languages:
Proficient: Python
Experienced: Bash, Nim, R, Rust,
Familiar: Awk, C, C#, C++, HTML, Java, Javascript, Matlab
Operating Systems: OS X, Linux, Windows

HONORS AND AWARDS
Best Poster Award 2018
Johns Hopkins University, CMDB Departmental Retreat

Thomas Hunt Morgan Fellowship 2015
Johns Hopkins University

PUBLICATIONS

Roach, N P, Sadowski N, Alessi AF, Timp W, Taylor J, and Kim JK. 2020. The full-length transcriptome of *C. elegans* using direct RNA sequencing. *Genome Research*
doi : https://doi.org/10.1101/gr.251314.119.

Fuller GG, Han T, Freeberg MA, Moresco JJ, Ghanbari Niaki A, **Roach, N P**, Yates 3rd JR, Myong S, and Kim JK. 2020. RNA promotes phase separation of glycolysis enzymes into yeast G bodies in hypoxia. *Elife* **9**.
doi : https://doi.org/10.7554/eLife.4848.

Jin M, Fuller GG, Han T, Yao Y, Alessi AF, Freeberg MA, **Roach, N P** , Moresco JJ, Karnovsky A, Baba M, et al.. 2017. Glycolytic enzymes coalesce in G bodies under hypoxic stress. *Cell Rep.* **20**: 895–908.
doi : https://doi.org/10.1016/j.celrep.2017.06.082.

Kapustina M, Tsygankov D, Zhao J, Wessler T, Yang X, Chen A, **Roach, N**, Elston TC, Wang Q, Jacobson K, et al.. 2016. Modeling the excess cell surface stored in a complex morphology of bleb-like protrusions. *PLOS Computational Biology* **12**: 1–25.
doi : https://doi.org/10.1371/journal.pcbi.1004841.

RELEVANT COURSEWORK

Bioinformatics

- Computational Genomics: Sequences
- Computational Genomics: Data Analysis
- Computational Genomics: Applied Comparative Genomics

Computation

- Machine Learning
- Algorithms & Analysis
- Data Structures
- Computer Organization
- Digital Logic & Computer Design
- Image Processing & Analysis

Biology

- Drug Discovery
- Advanced Cell Biology
- Graduate Biophysical Chem
- Genomes and Development
- Advanced Molecular Biology
- Cell & Developmental Biology
- Molecular Biology & Genetics
- Ecology & Evolution
- Mechanisms of the Cytoskeleton

RELEVANT COURSEWORK (CONTINUED)
Math

- Linear Algebra for Applications • Multivariate Calculus
- Discrete Math

CONFERENCE PRESENTATIONS

"Measuring the transcriptome of the *C. elegans* lifecycle using direct RNA sequencing." 2nd Annual London Calling Conference, Transcriptomics Breakout Session. May 2018.

"The full-length transcriptome of *C. elegans* using direct RNA sequencing" Genome Informatics, November 2019 (Poster)

INTERNAL PRESENTATIONS

"The full-length transcriptome of *C. elegans* using direct RNA sequencing" Johns Hopkins University CMDB departmental progress reports, November 2019.

"Interrogating the developmental transcriptome of *C. elegans* using direct RNA sequencing" Johns Hopkins University CMDB departmental progress reports, January 2019.

"Characterization of the C. elegans transcriptome by direct RNA sequencing" Johns Hopkins University Genomics Working Group Meeting. November 2018.

"Direct RNA Sequencing across *C. elegans* development" Johns Hopkins University CMDB Departmental Retreat, October 2018 (Poster).

"Building a developmental transcriptome of *C. elegans* using direct RNA sequencing" Johns Hopkins University CMDB departmental progress reports, April 2018.

"Nanopore sequencing of RNA: Computational Approaches and Initial Results" Johns Hopkins University CMDB Departmental Retreat, October 2017 (Poster).

INTERNAL PRESENTATIONS (CONTINUED)
"The Glycolytic Body: Characterizing a Novel Cellular Granule" Johns Hopkins University CMDB Departmental Retreat, October 2016 (Poster).

TEACHING EXPERIENCE

Johns Hopkins University *Teaching Assistant*

Quantitative Biology Lab Fall 2019

- Worked one-on-one with students to assist them in understanding the biological relevance behind, and computational processes necessary for their assignments

Quantitative Biology Bootcamp Fall 2019

- Worked hands on with students to ensure understanding of material
- Assessed student work to target areas for further discussion and explanation

Quantitative Biology and Biophysics Spring 2019

- Led weekly lab sessions, working directly with students to understand core concepts
- Developed an assignment focusing on simulating data and fitting Gaussian Mixture Models to data in Python

Human Genetics Fall 2018

- Presented a guest lecture on DNA sequencing

Introduction to the Human Brain Spring 2018
Chromosomes, Chromatin, and the Cell Nucleus Fall 2017
Protein Engineering and Biochemistry Lab Fall 2016 - Spring 2017